Functional Data Methods for the Analysis of Neuroimaging Data over Time

Sonia Dembowska

School of Mathematics

University of Leeds

A thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

September 2024

Intellectual Property

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Signed

Acknowledgements

I would like to thank my PhD supervisors, Professor Jeanine Houwing-Duistermaat and Dr Haiyan Liu, for their expertise and guidance during my PhD. Your advice, support, and confidence in me have enabled me to develop as a researcher.

I would like to acknowledge Professor John T. Kent, Dr Robert G. Aykroyd and Dr Georgios Aivaliotis for their comments at annual reviews, that have helped me in furthering my work. For fruitful academic discussions and support during parts of the project, I would like to thank Dr Nishant Ravikumar, who has been greatly generous with him time and energy to reply to my numerous queries.

The funding for my PhD has been provided by EPSRC. I would like to thank the Departments of Statistics at the University of Leeds for their work in bringing PhD student together.

I would like to thank my mom and my family and friends for their advice and continuous support. I'm incredibly grateful for everyone I have met along the way during my PhD that made this time more interesting and joyful. People such as Zandra Moore, Caroline Heide and Dr Olena Linnyk have broadened my perspective and supported me in different aspects of personal growth. Most importantly, I would like to thank my partner, Joe, who has stood with me this entire time and has heard me ramble on model assumptions, estimation methods and simulation designs so many times he could probably present my thesis himself.

Abstract

Neuroimaging has become ubiquitous in the study of human brain anatomy, yet it continues to pose multiple statistical challenges, namely high-dimensionality and spatial dependence. Functional data analysis (FDA) can address this by modelling observations as functions, applicable to images. In this thesis we consider neuroimaging datasets where a series of 3D images is captured over time, creating a 4 dimensional dataset for which there are no FDA methods. We model these data using FDA, proposing a novel model that preserves the spatial relations between voxels and simultaneously modelling the temporal correlation whilst maintaining computational efficiency.

Whilst all images are captured on a regular, high-dimensional grid, the time dimension can vary in density. This thesis considers two types of temporal data, densely collected images in the form of fMRI and sparse data collected longitudinally in a large-scale study. Current methods that model multi-dimensional functional data are limited to two dimensions and cannot be applied to imaging. We begin by introducing a non-parametric functional principal component model for the representation of spatio-temporal images as a product of time-invariant basis functions and subject specific score functions that can vary over time. We propose an estimation method that avoids calculating the covariance matrix, making our approach computationally efficient. The performance of the model and its estimation are studied via simulation. This method is applied to a task fMRI dataset. The obtained score functions are used to model the associations between brain activity and risk behaviour. In low dimensions we design a large simulation study to compare the performance of our model to state-of-the-art functional models. The simulation provides insight on appropriate use cases for the proposed model as well as shows that it has comparable performance in such cases.

The second type of functional data considered is on a sparse temporal grid. In this case we adapt our methodology to consider a longitudinal dataset of MRI images with missingness at several time points. In this application, the estimated score functions are modelled with a random slope model which described a subject's trajectory over time associated with a PC. The random intercepts and slopes are used to associate with, and later predict subject outcomes. Given that machine learning has become increasingly prevalent in image analysis for outcome prediction, we propose a deep neural network for disease state prediction. A framework for comparison between our proposed functional principal component (FPC) model and the network is proposed.

In this thesis, we propose a novel model for dimensionality reduction of images over time alongside an efficient estimation method. This method is investigated via simulation and is used to analyse two imaging datasets over dense and sparse temporal grids. Data analysis on fMRI and ADNI revealed that temporal variation plays an important role in outcome association and prediction. Whilst machine learning methods, especially neural networks, are frequently used in image analysis, FDA is particularly useful on small and complex datasets. Our approach provides an efficient and interpretable approach of modelling high dimensional data which can be used for association or prediction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Technological advance increased the availability of high-dimensional data, recorded continuously or intermittently over time. Many types of data can fall into this category, namely curves, images, shapes, or more general objects. The field of functional data analysis (FDA) provides tools and theory for analysis of such data, where each datum is regraded as a realization of a random function in the Hilbert space. Functional data was first discussed in Ramsay, 1982, and grew in popularity with several books being published on the topic making it widely accessible. Ramsay and Silverman, 2005 offers an approachable introduction to core mathematical and computational concepts of FDA whilst emphasizing its practical use on data. Ferraty and Romain, 2011 covers many of the core topics such as functional regression models and the functional principal components (FPC), but it also contains review papers with a focus on non-parametric methods and mathematical theory. Horvath and Kokoszka, 2012 covers the same core topics as Ferraty and Romain, 2011 but build on them with the construction of test statistics and the relevant asymptotic theory, with an emphasis on models for dependent functional data.

The first generation of functional data focused on a random sample of independent functions $X(t)$ on a bounded, compact interval $T \subset \mathbb{R}$. These functions are assumed to be square integrable, making them a stochastic process in the Hilbert space. Another frequent assumption is one of smoothness, which can help with regularization. Despite interest in understanding the underlying stochastic process and its properties, this process is rarely directly observable; data are often collected discretely, either on a fixed or random grid, which may be dense, sparse, or vary between subjects. Originally, this grid was assumed to be dense with regular spacing, which is exemplified by data such as signals from fMRI machines or curves from NMR spectroscopy. For many such cases, FDA provides considerable flexibility combined with natural ordering of data points on a domain that facilitate non or semi-parametric approaches.

In the last decade, a growing field of FDA focused on complex data objects, such as images or shapes (Wang et al., 2015). Simultaneously, medical and technical advances made imaging a frequently used tool in research, where it has become a common protocol in small and large multi-center longitudinal studies alike. Neuroimaging specifically has been used to further understand human psychological responses and the onset of neurological conditions. As part of this, a large number of publications began modelling neuroimaging data as functional objects. FDA on neuroimaging data provides an intuitive approach to modelling as if one treats the entire image as a function, the spatial correlations between voxels are preserved as opposed to traditional methods. Functional methods also allow for dimensionality reduction whilst maintaining spatial relations between voxels via functional principal components or smoothing techniques that can achieve more parsimonious representation of image data by using a basis expansion.

Images captured over time can be both found in the dense and sparse functional data categories. In this thesis, we are motivated by two such datasets. One, obtained on a dense temporal grid, is a set of fMRI images captured to understand the mechanisms behind risk in financial decisions (Mohr et al., 2010a). The other, captured on a sparse grid, is a set of T1 MRI images captured during a longitudinal study following a cohort of patients to understand the brain matter volume changes as a result of ageing and Alzheimer's Dementia (Mueller et al., 2005). Existing dimensionality reduction techniques for imaging, such as multivariate methods or machine learning, do not directly incorporate the spatial structure between voxels and across time points in the model, making them more suitable for large cross-sectional studies. They often require a large sample size, where the number of samples should be larger than the number of voxels in an image. Instead, FPC models can be estimated for cases where the number of images (n) is lower than the number of voxels (p). We aim to develop a dimensionality reduction method for images captured over time using the functional data analysis framework that can create summaries of such data as random variables over time which can be further used for association with or classification of outcomes.

Models for functional data over time exist in cases of low dimensions, where one dimensional curves are observed at several time points for subjects. These publications focused on the previously not considered case of functional data correlated over two domains, with the temporal domain being typically longitudinal. Initial models incorporated time spacing of the measurements into the coefficients through a linear structure, relying on additive assumptions (Greven et al., 2010). Later models introduced non-parametric models for time-correlated functional observations by assuming the covariance and mean functions to evolve smoothly over time (Chen and Müller, 2012; Park and Staicu, 2015). In traditional FDA approaches that do not include time-correlated functional data, images can be represented using b-splines and a covariance matrix can be estimated. However, in the case of these models such a representation has a non-trivial solution and hence neither of these models can be directly implemented on neuroimaging data, where the covariance matrix would be 6 dimensional and for images over time it would have an additional 2 dimensions.

Few publications in the field of functional data analysis model images over time. The ones that do exist often consider the dense temporal case and model fMRI images as a time-series with a semi-parametric model (van Bömmel et al., 2013; Park et al., 2009). In this thesis, we consider a non-parametric approach to modelling high dimensional, functional data correlated over space and time. We focus on the dimensionality reduction of images into a linear product of time-invariant basis function varying over space s and subject specific score function over t. This decomposition allows for a parsimonious representation of images that can be further used to associate or predict outcomes via score functions.

Linear and non-linear dimensionality reduction methods that could be used alternatively on such data do not consider the temporal correlation between observations. Additionally, they often require sample sizes larger than the number of voxels in an image and, with the case of non-linear methods, may lack interpretability. Our methods provide an explainable and intuitive approach using the FDA framework that can be applied to small datasets and has a computationally efficient estimation. We provide several improvements on existing literature within FDA. Firstly, we provide a non-parametric approach for the dimensionality reduction of imaging data, where complete 3D images are used. Secondly, the model can be estimated efficiently by avoiding the estimation of the covariance matrix. Finally, our model can be used for association whilst existing ML methods can primarily be use in prediction or classification.

Our work makes contributions in both statistical methodology and applications within the field of FDA, specifically applied to neuroimaging datasets over time. In particular we make the following contributions:

1. Current available methodology (Park and Staicu, 2015) can only model curves as a sum of products of basis functions and time-varying scores in low dimensions. For the representation of 3D imaging data over time, we developed a functional principal component model as a product of time-invariant spatial principal component functions and subject specific score functions over time.

2. A time and computationally efficient estimation method for fitting this model was developed and implemented in R, avoiding the direct computation of the covariance matrix and its performance was investigated via simulation.

3. In low dimensions, where other methods are available, we design a simulation to compare the performance of our model to existing models as well as explicitly show the computational advantage of our estimation method. The simulation varies in designs, sample size and data dimensionality to account for multiple scenarios.

4. With our novel model, we studied the associations between brain regions and patient outcomes in two datasets, one including dense time observations (fMRI) in a small number of subjects and one including sparse time observations (longitudinal) in a larger number of subjects. For the first dataset, we studied the association between active brain regions and subject risk-averseness. For the second dataset, we found

associations between brain regions and Alzheimer's disease.

5. We propose a framework to compare the prediction performance and feasibility of our functional model to a neural network that takes in full 3D images and is trained to learn subject specific trajectories over time.

1.2 Structure of the Thesis

This thesis has eight chapters. This section provides an overview of how the chapters are organized. The current is Chapter 1, which is a general introduction of the thesis.

Chapter 2 will provide an introduction to functional data analysis from the formulation of functional data to functional principal component analysis and functional regression. It includes all the elements that are necessary for the understanding of our contributions in the field.

Chapter 3 introduces neuroimaging as well as common pre-processing steps that take place prior to image analysis. We then introduce the two datasets used in later chapters of the thesis as well as relevant previous analysis approaches. These datasets are a task fMRI study to understand the brain regions responsible to risk decision making and a longitudinal MRI dataset from the ADNI study, following elderly patients to better understand the onset and progression of Alzheimer's dementia.

Chapter 4 introduces a novel model using the FDA framework. We define a new functional principal component model and an efficient estimation method that circumvents calculating the high dimensional covariance matrix. Our model represents data using spatial principal components whilst the score will contain temporal information. Our novel contributions are as follows: firstly, we propose an algorithm capable of representing high-dimensional datasets captured in space and time. Secondly, the estimation algorithm is computationally efficient by using singular value decomposition. Thirdly, we will demonstrate that our parameter estimation method is accurate with a simulation study. Finally, the data analysis will recover active brain regions and associate their activity over time with subject's risk attitude.

Chapter 5 focuses on the comparison of the model introduced in Chapter 4 to state of the art in literature. The simulation is done in lower dimensions to allow for this comparison. Additionally, we investigate the effect of sample size, and data dimensionality on computation time required for model estimation. The simulation varies in design complexity, number of samples and noise. We discuss the results as well as show the limitations of our method within certain specific scenarios.

In Chapter 6, we extend our proposed model from Chapter 4 to sparse, longitudinal imaging data. First, we study the effect of missingness in a simulation study, then we apply the model to images from the ADNI dataset to model associations between regions of the brain and the presence of dementia. The scores obtained for each subject are assumed to follow a linear mixed model where each subject has a random intercept and a random slope which can represent the subject's temporal trajectory and can be used in a logistic regression to associate with outcomes. Finally, we investigate model fit on different subsets of the data.

Chapter 7 aims to compare our proposed model to a neural network, given that deep learning has increased in popularity for image analysis. We adapt a published network to fit our data. A framework is proposed for the comparison of networks by performing cross validation to evaluate our methods in terms of outcome prediction.

The final chapter gives our conclusions, future work and possible extensions of the thesis.

Chapter 2

Introduction to Functional Data Analysis

2.1 Introduction

Functional data analysis (FDA) considers each datum as a realisation of an infinite dimensional function defined over some set T, with leading publications being Ramsay and Silverman, 2005; Horvath and Kokoszka, 2012; Hsing and Eubank, 2015; Ferraty and Romain, 2011. Chapter 7 of Hsing and Eubank, 2015 highlights the two different perspectives on functional data. One considers functional data to be realizations of random variables taken in the Hilbert space whereas the other considers functional data as sample paths of a stochastic process with smooth mean and covariance functions. Both perspectives provide the theoretical background of concepts like mean and covariance, as well as the foundations for the tools used to analyse the variability of a sample. In this chapter, we take the former approach and treat functional data as functions in the Hilbert space. We introduce the mathematical concepts underlying FDA: smoothing, functional principal component analysis and functional regression. This section is primarily influenced by Horvath and Kokoszka, 2012; Ramsay and Silverman, 2005, and Hsing and Eubank, 2015.

2.2 Foundations of Functional Data Analysis

2.2.1 The Space L^2

The space $L^2[\mathcal{T}]$ is a separable Hilbert space of measurable real-valued functions $f(\cdot)$ for which the integral of the square is finite i.e

$$\int f^2(t)dt < \infty. \tag{2.1}$$

It is often assumed that $\mathcal{T} \subset \mathbb{R}^d$, and, without loss of generality, we can assume $\mathcal{T} = [0, 1]$. The space is endowed with the inner product

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

and the norm

$$||f|| = \big(\int f^2(t)dt\big)^{\frac{1}{2}}, \tag{2.3}$$

for any $f, g \in L^2[\mathcal{T}]$. If $f, g \in L^2$, then $f = g$ means $\int \big[f(t) - g(t)\big]^2 dt = 0$. Let $\mathcal{L}^2$ denote the space of bounded operators (transformations on vector spaces) on elements of L^2, consider an operator $\Psi[f(t)] \to g(t)$ for $f, g \in L^2$, with the norm

$$||\Psi||_{\mathcal{L}} = \sup\{||\Psi[f(\cdot)]|| : ||f|| \leq 1\}. \tag{2.4}$$

An operator Ψ is said to be compact if there exist two orthonormal bases $\{a_j(\cdot)\}$ and $\{b_j(\cdot)\}$ and a real, non-negative sequence $\{\lambda_j\}$ converging to zero such that

$$\Psi[f(\cdot)] = \sum_{j=1}^{\infty} \lambda_j \langle f, a_j\rangle b_j, \text{ for } f(\cdot) \in L^2. \tag{2.5}$$

In other terms, it could be said that every compact operator is an operator that can be represented using singular value decomposition (SVD). The orthonormal bases $\{a_j(\cdot)\}$ and $\{b_j(\cdot)\}$ are analogous to the unitary matrices in ordinary multivariate SVD and $\{\lambda_j\}$ would be the entries of the diagonal matrix.

An operator is said to be symmetric if, for $f(\cdot), g(\cdot) \in L^2[\mathcal{T}]$,

$$\langle \Psi[f], g \rangle = \langle f, \Psi[g] \rangle, \tag{2.6}$$

positive-definite if $\langle \Psi[f], g \rangle > 0$ and positive semidefinite if $\langle \Psi[f], g \rangle \geq 0$.

A compact operator with the real sequence $\{\lambda_j\}$ satisfying $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$ is a Hilbert-Schmidt operator. A symmetric, positive semidefinite Hilbert-Schmidt operator admits the decomposition from eq. 2.5 with $\{a_j(\cdot)\} = \{b_j(\cdot)\}$ and $\{a_j\}$ is the set of eigenfunctions satisfying: $\Psi[a_j] = \lambda_j a_j$.

Consider a bounded linear operator $\Psi \in \mathcal{L}^2$, defined as:

$$\Psi[f(t)] = \int \psi(t,s) f(s) ds, \qquad f \in L^2, \tag{2.7}$$

with the real kernel $\psi(\cdot,\cdot)$, that is a bivariate function in $L^2[\mathcal{T} \times \mathcal{T}]$. Ψ is said to be Hilbert-Schmidt if and only if

$$\int \int \psi^2(s,t) ds dt < \infty, \tag{2.8}$$

which leads to

$$||\Psi||_{\mathcal{L}}^2 = \int \int \psi^2(s,t) ds dt. \tag{2.9}$$

If $\psi(s,t) = \psi(t,s)$ and for any $f(\cdot) \in L^2[\mathcal{T}]$, $\int \int \psi(s,t) f(s) f(t) ds dt \geq 0$ then the integral operator Ψ is symmetric and semi-positive definite. Any symmetric and positive semi definite operator in L^2 has non-negative eigenvalues. This allows us to state the following Mercer's Theorem.

Theorem 1 (Mercer's Theorem)**.** *Any symmetric and semi-positive definite operator* Ψ *with a kernel* $\psi(\cdot,\cdot)$ *forms an orthonormal basis* $\{v_j(\cdot)\}_l \in L^2$ *that solves the eigendecomposition problem:*

$$\Psi[v_j(s)] = \int \psi(s,t) v_j(t) dt = \lambda_j v_j(s). \tag{2.10}$$

Then, its kernel $\psi(s,t)$ has the representation (Aizerman et al., 1964):

$$\psi(s,t) = \sum_{j=1}^{\infty} \lambda_j v_j(t) v_j(s), \tag{2.11}$$

where λ_j are positive eigenvalues of Ψ in decreasing order ($\lambda_j \geq \lambda_{j+1}$) and v_j are the eigenvectors of Ψ.

Mercer's theorem (Mercer, 1909; Riesz and Sz.-Nagy, 1956) follows from eq.(2.5). A complete proof can be found in Ghojogh et al., 2021 and an illustration can be found below.

Consider a Hilbert Schmidt operator Ψ defined by equation (2.7), where its kernel is $\psi(s,t)$, a symmetric and positive definite function in $L^2[\mathcal{T}]$. Ψ is semi positive definite and admits the decomposition from equation (2.5) with basis vectors satisfying the equation

$$\Psi[v_j] = \lambda_j v_j. \tag{2.12}$$

We can then use it to show that:

$$\lambda_j \langle v_j, v_j \rangle = \lambda_j \int v_j(s) v_j(s) ds = \int \lambda_j v_j(s) v_j(s) ds = \int \Psi v_j(s) v_j(s) ds \geq 0, \tag{2.13}$$

as $\langle \cdot, \cdot \rangle$ is positive definite, $\lambda_j \geq 0$. Define $\varphi_j(t) \in L^2[\mathcal{T}]$ by $t \mapsto \sqrt{\lambda_j} v_j(t)$. Then

$$\begin{aligned} \sum_{j=1}^{\infty} \int \langle \varphi_j(t), \varphi(s) \rangle f(s) ds &= \int \sum_{j=1}^{\infty} \lambda_j v_j(t) v_j(s) f(s) ds & (2.14) \\ &= \sum_{j=1}^{\infty} \lambda_j v_j(t) \int u_j(s) f(s) ds & (2.15) \\ &= \sum_{j=1}^{\infty} \lambda_j \langle f, v_j \rangle v_j(t) = \Psi f(t). & (2.16) \end{aligned}$$

Since the above holds for all $f \in L^2$ then:

$$\psi(s,t) = \langle \varphi_j(t), \varphi(s) \rangle = \sum_{j=1}^{\infty} \lambda_j v_j(t) v_j(s). \tag{2.17}$$

2.2.2 Random functions in L^2

Let $X(t)$ be a random function in $L^2[\mathcal{T}]$. We say that X is integrable if

$$\mathbb{E}||X|| = \mathbb{E}\big(\int X^2(t)dt\big)^{\frac{1}{2}} < \infty. \tag{2.18}$$

If $\mathbb{E}||X|| < \infty$, then there exists a unique function $\mu \in L^2$ such that $\mathbb{E}\langle y, X\rangle = \langle y, \mu\rangle$ for any $\mu \in L^2$, which implies that $\mu = \mathbb{E}[X(t)]$. The expectation commutes with bounded operators so for all bounded and continuous linear operators $\Psi \in \mathcal{L}$, we have $\mathbb{E}[\Psi(X)] = \Psi(\mathbb{E}[X])$.

If X is square integrable, i.e $\mathbb{E}||X||^2 < \infty$, then the second order variations of X are encoded in the covariance function:

$$c(s,t) = \mathbb{E}[(X(s) - \mu(s))(X(t) - \mu(t))], \quad s,t \in \mathcal{T}. \tag{2.19}$$

The variance function is the case of $c(s,t)$ where $s = t$. For simplicity, assume that $\mathbb{E}[X] = 0$, and consider the covariance operator $C : L^2[\mathcal{T}] \to L^2[\mathcal{T}]$ defined as:

$$C[f(t)] = \mathbb{E}\big[\langle X, f\rangle X\big] \text{ for } f \in L^2, \tag{2.20}$$

which can be rewritten to include the kernel function $c(\cdot,\cdot)$:

$$Cf(t) = \int c(t,s)f(s)ds. \tag{2.21}$$

The covariance function satisfies two properties: $c(t,s) = c(s,t)$ as the expectation is commutative, and

$$\begin{aligned}\int\int c(t,s)f(s)f(t)dsdt &= \int\int \mathbb{E}\big[X(t), X(s)]f(s)f(t)dsdt \\ &= \mathbb{E}\Big[\big(\int X(t)f(t)dt\big)^2\Big] \geq 0.\end{aligned}$$

Hence the covariance operator C is symmetric and semi-positive definite. Therefore, it has non-negative eigenvalues denoted as λ_j and eigenvectors v_j that satisfy the equation $C[v_j] = \lambda_j v_j$. The functions v_j are orthogonal and can be normalised to have unit norm so that they form an orthonormal basis in L^2. Given $\{v_j\}$, we can use Parseval's identity

$$||X||^2 = \sum_{j=1}^{\infty} |\alpha_j|^2 ||v_j||^2 = \sum_{j=1}^{\infty} |\alpha_j|^2 \tag{2.22}$$

where $\alpha_j = \langle X, v_j \rangle / \langle v_j, v_j \rangle$ are the coefficients of X in the system $\{v_j\}$. One can see that $\mathbb{E}[\alpha_j] = \lambda_j$ and thus we have:

$$\sum_{j=1}^{\infty} \lambda_j = \sum_{j=1}^{\infty} \mathbb{E}\Big[\big\langle X, v_j \big\rangle^2\Big] = \mathbb{E}||X||^2 < \infty. \tag{2.23}$$

A function that satisfies the above properties, meaning it is symmetric, semi-positive definite and satisfies eq. (2.23) is a proper covariance function.

By Mercer's Theorem, the covariance function has the form:

$$c(s,t) = \sum_{j=1}^{\infty} \lambda_j v_j(t) v_j(s), \tag{2.24}$$

with λ_j and $\{v_j\}$ eigenvalues and eigenvectors corresponding to the covariance operator C.

Set the covariance function $c(s,t)$ as the Mercer's kernel to the covariance operator $CX(t)$ and let $v_j(t)$ be an orthonormal basis in L^2 as in eq. (2.24). Then $X(t)$ admits the Karhunen Loéve expansion:

$$X(t) = \sum_{k}^{\infty} \psi_k v_k(t), \tag{2.25}$$

where $\psi_k = \int X(t) v_k(t) dt$ are random uncorrelated weights with zero-mean and variance λ_k $\mathbb{E}[\psi_k] = 0, \forall k \in \mathbb{N}$ and $\mathbb{E}[\psi_i \psi_j] = \delta_{ij}\lambda_j, \forall i,j \in \mathbb{N}$ with δ_{ij} denoting the Kronecker delta.

The Karhunen Loéve expansion converges uniformly in L^2. To prove this, let $X_K(t) =$

$\sum_k^K \psi_k v_k(t)$ for some integer K, then

$$\begin{aligned}\mathbb{E}\Big[|X(t) - X_K(t)|^2\Big] &= \mathbb{E}[X^2(t)] + \mathbb{E}[X_K^2(t)] - 2\mathbb{E}[X(t)X_K^2(t)] \\ &= c(t,t) + \mathbb{E}\Big[\sum_k^K\sum_l^K \psi_k\psi_l v_k(t)v_l(t)\Big] - 2\mathbb{E}\Big[X(t)\sum_k^K \psi_k v_k(t)\Big] \\ &= c(t,t) + \sum_k^K \lambda_k v_k^2(t) - 2\mathbb{E}\Big[\sum_k^K \int X(t)X(s)v_k(s)v_k(t)ds\Big] \\ &= c(t,t) - \sum_k^K \lambda_k v_k^2(t)\end{aligned}$$

which converges by Mercer's Theorem.

2.2.3 Estimators and Assumptions

In practice we observe a set of n curves $\{X_i\}_{i=1,\ldots,n}$ where each curve is a realisation of a random function $X(\cdot) \in L^2$. The key assumptions made on the random variables are:

AS 1. *X are independently and identically distributed (iid.),*

AS 2. $\mathbb{E}[X] = 0$,

AS 3. $\mathbb{E}||X||^4 < \infty$.

These assumptions ensure the convergence of the sample mean $\hat{\mu}(t)$, the sample covariance function $\hat{c}(s,t)$ and the sample covariance operator $\hat{C} : \{L^2\} \to \{L^2\}$ defined as:

$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^n X_i(t), \tag{2.26}$$

$$\hat{c}(s,t) = \frac{1}{n}\sum_{i=1}^n \big(X_i(s) - \hat{\mu}(s)\big)\big(X_i(t) - \hat{\mu}(t)\big), \tag{2.27}$$

$$\hat{C}\big(f(s)\big) = \frac{1}{n}\sum_{i=1}^n \Big\langle X_i(t) - \hat{\mu}(t), f(t)\Big\rangle\Big(X_i(s) - \hat{\mu}(s)\Big), \text{ for } f(\cdot) \in L^2. \tag{2.28}$$

If Assumption AS (1) holds, then the sample mean function exists and is unique. Assumptions AS (2, 3) assure the convergence of the estimated covariance operator to the true one.

Let $p \in \mathbb{Z}$ denote the number of eigenvalues and eigenfunctions given by the solution to the equation $\hat{C}[v_j] = \lambda_j v_j$, with $\hat{C}$ denoting the estimator of the covariance operator. Typically, the eigenfunctions v_j are normalized so that $||v_j|| = 1$. We assume that eigenvalues are in strictly decreasing order, i.e $\lambda_1 \geq \cdots \geq \lambda_p > 0$, making them identifiable. The solution to this equation are the estimators for eigenvalues and eigenfunctions of the population covariance operator, denoted as $\hat{\lambda}_j$ and $\hat{v}_j(s)$ for $j \in \{1, \ldots, p\}$. The eigenfunctions $\hat{v}_j(s)$ can be rescaled to have a unit norm. This ensures the functions form a unique solution to the eigendecomposition problem. Both $\hat{\lambda}_j$ and $\hat{v}_j(s)$ are shown to be consistent and unbiased estimators and $\hat{v}_j(s)$ are identifiable up to a sign. By Mercer's Theorem, the estimators of the eigenelements solve:

$$\int \hat{c}(s,t)\hat{v}_j(s)ds = \hat{\lambda}_j \hat{v}_j(t) \tag{2.29}$$

for $j = 1, \ldots, p$.

2.3 Functional Principal Component Analysis

Functional Principal Component Analysis is a tool for the reduction of an infinite dimensional functional object to a finite one. It can be introduced in two ways, as a set of basis vectors that maximise data variability or as an optimal orthonormal basis to reduce the error between the true functional object and the one with reduced dimensions. For a random function $X(t) \in L^2[\mathcal{T}]$, we want to find an orthonormal basis $\{u_j\}_j^p$, for a fixed integer p, that minimizes the expression:

$$S^2 = \mathbb{E}\Big|\Big|X_i(t) - \sum_j^p \langle X_i, u_j\rangle u_j(t)\Big|\Big|^2 \tag{2.30}$$

This is analogous to maximising

$$Var = \mathbb{E}\Big|\Big|\sum_j^p \langle X_i, u_j\rangle u_j(t)\Big|\Big|^2. \tag{2.31}$$

Given realizations $X_i(t), i \in \{1, \ldots, N\}$ of a random function $X(t) \in L^2[\mathcal{T}]$, S^2 is approximated by:

$$\hat{S}^2 = \sum_{i=1}^N \Big|\Big|X_i - \sum_{j=1}^p \langle X_i, u_j\rangle u_j\Big|\Big|^2 \tag{2.32}$$

by finding an optimal basis u_j. If such a basis is found then we can replace the complete curve $X_i(t)$ with its approximation $\sum_{j=1}^{p}\langle X_i, u_j\rangle u_j$. This set of functions is called the optimal empirical orthonormal basis.

The functions $\{u_j\}_{j=1}^{p}$ minimizing $\hat{S}^2$ are equal (up to a sign) to the normalized eigenfunctions of the sample covariance operator (Bosq, 2000; Dauxois et al., 1982). Horvath and Kokoszka, 2012 show this in Chapter 3 by setting $p = 1$. We want to find u s.t. $||u|| = 1$ which minimises:

$$\begin{aligned}\sum_{i=1}^{N}\Big|\Big|X_i - \langle X_i, u\rangle u\Big|\Big|^2 &= \sum_{i=1}^{N}||X_i||^2 - 2\sum_{i=1}^{N}\langle X_i, u\rangle^2 + \sum_{i=1}^{N}\langle X_i, u\rangle^2||u||^2 \\ &= \sum_{i=1}^{N}||X_i||^2 - \sum_{i=1}^{N}\langle X_i, u\rangle^2\end{aligned}$$

As this solution is unique, we conclude that $u = \hat{v_1}$. The same reasoning is applied for the case where $p > 1$.

Therefore, the basis that minimises $\hat{S}^2$ is formed of the eigenfunctions of the covariance operator. Thus, if we wish to use the representation of a random function $X(t)$ as a product of an orthonormal basis satisfying eq. (2.30), we can apply the Karhunen Loéve Theorem set with the covariance function as the Mercer's kernel. This yields the result:

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty}\psi_{ij}v_j(t), \tag{2.33}$$

where ψ_{ij} are the principal component scores equal to $\langle X_i(t), v_j(t)\rangle$. It is also easy to show that, given the relation $Cv_j = \lambda_j v_j$:

$$Var(\psi_{ij}) = \mathbb{E}[\langle X_i(t), v_j(t)\rangle^2] = \langle[\langle X_i, v_j\rangle X], v_j\rangle = \langle Cv_j, v_j\rangle = \lambda_j. \tag{2.34}$$

A similar approach can be utilised for the population variance:

$$\mathbb{E}[X(t)^2] = \sum_{j=1}^{\infty}\langle[\langle X, v_j\rangle X], v_j\rangle = \sum_{j=1}^{\infty}\langle Cv_j, v_j\rangle = \sum_{j=1}^{\infty}\lambda_j. \tag{2.35}$$

Naturally, we cannot compute an infinite number of principal components so the summation is truncated after p components. The most frequent way of determining the number

p is by the cumulative percentage of total variance (CPV) explained by the first p PCs as a proportion over all N available PCs:

$$CPV(p) = \frac{\sum_{j=1}^{p} \lambda_j}{\sum_{j=1}^{N} \lambda_j}$$

2.4 Smoothing

In FDA, smoothing is applied to the observed discrete data to obtain a noise-free estimate of the underlying function $X(t)$ that will be used in further analysis. In a large data setting, smoothing can allow for a simpler representation of the raw data. Ramsay and Silverman, 2005 provide a good overview of useful smoothing techniques that include basis expansion or derivatives. For multi-dimensional settings, one can use tools such as kernel or sandwich smoothers.

A common smoothing approach would be to represent a function by basis expansion. A basis function system is a set of known functions $\{\phi_k\}$ that are mathematically independent of each other. It is defined such that one can approximate any function arbitrarily well by taking a weighted sum or a linear combination of a sufficiently large number K of these functions (Ramsay and Silverman, 2005). These functions can take the form of monomials, Fourier series, or more modern techniques such as b-splines and wavelets. The choice of type and number of basis functions depends on the underlying raw data structure and can affect the reconstruction of the original data as well as the results of further analysis. Here we will go over Fourier and b-spline vectors, as they are used further in the thesis.

Fourier basis functions

The Fourier basis is a periodic basis defined by the parameter ω defining the period $2\pi/\omega$. The basis functions $\{\phi_k\}_K$ are defined iteratively, where for any integer $r < K/2$:

$$\phi_0 = 1, \quad \phi_{2r-1} = \sin(r\omega t), \quad \phi_{2r} = \cos(r\omega t)$$

If the observations on $t \in \mathcal{T}$ are equally spaced and the period is equal to the length of interval $\mathcal{T}$, then the basis is orthogonal.

A Fourier basis is useful with functions stable over the whole domain with no strong local features and where the curvature seems to be of the same order everywhere. It is preferable that periodicity is also present in the raw data itself, examples of periodic data include annual rainfall or gait data. Fourier series generally yield expansions which are uniformly smooth. Properties and derivatives of the Fourier series are well known and the fast Fourier transform makes the calculation of coefficients efficient.

Despite its advantages, this basis expansion could be inappropriate in cases where the underlying data may have discontinuities. Ramsay and Silverman, 2005 summarise it quite humorously: "a Fourier series is like margarine: it's cheap and you can spread it on practically anything, but don't expect that the result will be exciting eating".

B-splines

Splines are the most common choice for approximation of non-periodic functions as they provide fast computation of polynomial fitting and have high flexibility. Whilst, there are several splines that were introduced, the most common and the one that is used throughout this thesis are b-splines. (**DeBoor1978ASplines**)

Splines are piecewise polynomial functions $B_l^q(t)$ defined on some domain $\mathcal{T}$ of degree q which are joined together at some points called knots, an increasing sequence of points $\tau_1 < \tau_2 < \cdots < \tau_L$ on the domain. The domain $\mathcal{T}$ is divided into $L-1$ equal intervals by L knots, where $L \geq q+2$. Each interval will be covered by $q+1$ b-splines of degree q. As the space of b-splines is a vector space, any linear combination of splines are again splines. Eilers and Marx, 1996 provide general properties of splines for b-splines of degree q:

- it consists of $q+1$ polynomial pieces each of degree q,
- the polynomial pieces join at q inner knots,
- at the joining points, the derivatives up to the order of $q-1$ are continuous,
- the b-spline is positive on a domain $\mathcal{T}$, everywhere else it is zero,

- at the boundaries, it overlaps with $2q$ polynomial pieces of its neighbours
- at a given t, $q+1$ b-splines are non-zero.

B-splines can be constructed using the Cox–de Boor recursion formula. Given a set of L knots, it starts with the B-splines of degree $q=0$, i.e a piecewise constant polynomial: $B_l^0 = 1$ if $\tau_l \leq t < \tau_{l+1}$ and 0 otherwise. The higher order b-spline basis of degree $q > 0$ is given by the functions $B_l^q(t)$ for $l \in \{1, \dots, L-1\}$, defined recursively:

$$B_l^q(t) = \frac{t - \tau_{l-q}}{\tau_l - \tau_{l-q}} B_{l-1}^{q-1}(t) + \frac{\tau_{l+1} - t}{\tau_{l+1} - \tau_{l+1-q}} B_j^{q-1}(t). \tag{2.36}$$

B-spline basis functions have useful properties such as a compact support (see point 3 of general properties) and, for any point within the domain $[\tau_1, \tau_l]$, the basis functions sum to 1. 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.

The choice of degree q and the number of knots L are dependant on the application. Typically the choice of degree is more restricted and most commonly natural splines are used. Knots are most often spaced out evenly, a finer grid may allow for more detailed data representation but it also increases the degrees of freedom. One way of dealing with this is through applying roughness penalties in order to find a balance between bias and the variance of fit (Chapter 4.5 of Ramsay and Silverman, 2005).

Fitting basis functions

Given observations $Y_i(t_j)$ for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, J\}$ where $t_j \in \mathcal{T}$ for all j, we can represent them using the model:

$$Y_i(t_j) = X_i(t_j) + \varepsilon_i(t_j), \tag{2.37}$$

where $\varepsilon_i(t_j)$ is the error term generally assumed to be normally distributed with mean zero and with function covariance equivalent to $Cov(Y_i(t_j), Y_i(t_{j'}))$ for $t_j, t_{j'} \in \mathcal{T}$. We can then use a basis expansion for a set of basis vectors $\{\phi(t)\}_k$:

$$X_i(t_j) = \sum_k^K c_{ik}\phi_k(t_j). \tag{2.38}$$

The coefficients c_{ik} can be estimated using ordinary or weighted least squares.

2.5 Functional Regression

Functional linear models, like their multivariate counterparts, are useful in a broad range of applications and hence, are a heavily researched topic of FDA. There are three cases of regression, in each one, the functional data takes on a different role in model: either as a response variable or a regressor or as both. For simplicity we assume that the responses and the covariates have mean zero and that the errors ε_i are independent of the explanatory variables X_i.

The first is a functional response model or function-on-scalar, where which the responses are curves, but the regressors are known scalars. It can be expressed as:

$$Y_i(t) = \beta(t)X_i + \varepsilon_i(t), \tag{2.39}$$

with $\varepsilon_i(t)$ assumed to be a normally distributed with mean zero and with function covariance $Cov(Y_i(t_j), Y_i(t_{j'}))$ for $t_j, t_{j'} \in \mathcal{T}$.

The scalar response model or scalar-on-function regression, where at least one of the regressors is functional and the response is scalar. It is written in the form:

$$Y_i = \int \beta(s)X_i(s)ds + \varepsilon_i \tag{2.40}$$

where $\varepsilon_i \sim N(0, \sigma^2)$.

The third type is the fully functional (function-on-function) model where the response and predictor are functional. This takes the form

$$Y_i(t) = \int \beta(t,s)X_i(s)ds + \varepsilon_i(t) \tag{2.41}$$

where $\varepsilon_i(t)$ defined as before.

In the cases where the mean of the response variable is non zero an additional term is added that acts as an intercept, if the outcome is scalar then it is a coefficient α and it the outcome is functional then it will be of the form $\beta_0(t)$.

It is useful to express both the functional regressor and coefficients as a linear combination of basis functions $\{\phi(\cdot)\}_k^{K_X}$ and $\{\varphi(\cdot)\}_k^{K_\beta}$:

$$X_i(s) = \sum_k^{K_X} x_{ik}\phi_k(s) \tag{2.42}$$

$$\beta(s) = \sum_k^{K_\beta} b_k\varphi_k(s)$$

Consider the $N \times K_X$ matrix x where each row is the vector $x_i = [x_{i1}, \ldots, x_{iK_X}]$, and vector $b = [b_1, \ldots, b_{K_\beta}]^T$ for the coefficients. The basis functions will be stored in matrices $\phi(t) = [\phi_1(t), \ldots, \phi_{K_X}(t)]^T$, $\varphi(t) = [\varphi_1(t), \ldots, \varphi_{K_\beta}(t)]^T$. The response can be formulated to be a vector of length N of functions $Y_i(t), i \in \{1, \ldots, N\}$ or scalars Y_i, and will be denoted as $Y(t)$ and Y, respectively. These matrix representations can be used to simplify the regression equations.

For scalar-on-function regression, equation (2.40) would take the form:

$$\begin{aligned} Y &= \int \beta(s)X(s)ds + \varepsilon \\ &= \int \Big[\sum_k^{K_\beta} b_k\varphi_k(s)\Big]\Big[\sum_k^{K_X} x_k\phi_k(s)\Big]ds + \varepsilon \\ &= x^T\Big[\int \varphi_k(s)\phi_k(s)ds\Big]b + \varepsilon \\ &= x^T W_{\phi\varphi} b + \varepsilon. \end{aligned}$$

One can define a matrix $Z = [x^T W_{\phi\varphi}]$ and then the model becomes $\hat{Y} = Zb$. Thus the estimator for $\beta(\cdot)$ is defined as:

$$\hat{\beta}(t) = \sum_k^{K_\beta} \hat{b}_k\varphi_k(s) \tag{2.43}$$

where $\hat{b} = (Z^T Z)^{-1} Z^T Y$. Further approaches using the roughness penalty as a form of regularization are discussed in section 15.4 of Ramsay and Silverman, 2005 or section 8.4 of Horvath and Kokoszka, 2012.

In the case of function-on-function regression, the response can be additionally represented as a linear combination of coefficients y_{il} and basis functions $\{\chi(t)\}_l^{K_Y}$ that can be stored as a matrix $\chi(t) = [\chi_1(t), \ldots, \chi_{K_Y}(t)]^T$. In this case, the functional coefficient $\beta(t,s)$, using $\{\chi(t)\}_l^{K_Y}$ and $\{\phi(s)\}_k^{K_X}$, can be expressed as a linear combination of the basis vectors:

$$\beta(t,s) = \sum_k^{K_Y} \sum_l^{K_X} b_{kl} \chi_k(t) \phi_l(s) \tag{2.44}$$

$$= \phi^T(s) B \chi(t), \tag{2.45}$$

where B is a $K_Y \times K_X$ matrix of coefficients. Chapter 8 of Horvath and Kokoszka, 2012 uses these representations to achieve an estimate for $\beta(t,s)$. Let Z^* be a $N \times K_X$ matrix defined as $Z^* = \int X(s)\phi^T(s)ds$. Then equation (2.41) can be expressed using matrix notation:

$$Y(t) = Z^* B \phi(s) + \varepsilon(t) \tag{2.46}$$

If we define a $K_Y \times K_Y$ matrix $J = \int \chi(t)\chi^T(t)dt$, then we have:

$$\int Y(t)\chi^T(t)dt = Z^* B J + \int \varepsilon(t)\chi^T(t)dt. \tag{2.47}$$

Multiplying by Z^{*T} and ignoring the error term gives:

$$Z^{*T} \int Y(t)\chi^T(t)dt = Z^{*T} Z^* B J. \tag{2.48}$$

To solve for B, we rewrite this using the kronecker product $\otimes$:

$$(J^T \otimes [Z^{*T} Z^*]) vec(B) = vec\Big(Z^{*T} \int Y(t)\chi^T(t)dt\Big). \tag{2.49}$$

then, if J and $Z^{*T}Z^{*}$ are non-singular, a unique solution exists

$$vec(B) = (J^T \otimes [Z^{*T}Z^{*}])^{-1} vec\Big(Z^{*T} \int Y(t)\chi^T(t)dt\Big). \qquad (2.50)$$

An alternative approach to the estimation of $\beta(t,s)$ discussed in Ramsay and Silverman, 2005, Chapter 16. This approach allows for large K_X, K_Y and introduce a roughness penalty.

The function-on-scalar regression equation (2.39) can be generalised to includes multiple (p) covariates stored in a matrix X of dimension $N \times p$. The coefficient function is fitted to minimise the least squares criterion:

$$LS(\beta) = \int \big[Y(t) - \beta(t)X\big]^T \big[Y(t) - \beta(t)X\big] dt. \qquad (2.51)$$

The estimator for $\beta(t)$ can be obtained by using a similar approach to the function-on-function regression. Different approaches are discussed in section 12.4 of Ramsay and Silverman, 2005.

When $X_i(s)$ and $Y_i(t)$ have basis expansions: $\{\phi(s)\}_l^{K_X}$ and $\{\chi(t)\}_l^{K_Y}$. The coefficient $B(t,s)$ simplifies to equation (2.45). In this setting, FPCA for both $X_i(s)$ and $Y_i(t)$ has been used independently to reduce the dimensionality of the functional data Yao et al., 2005. Functional partial least squares approaches are another way of functional regression that takes into account the joint variability of the outcome and regressors (Preda and Saporta, 2005; Preda et al., 2007).

2.6 Temporal and multidimensional FDA

Much of the early work in FDA was concerned 1-dimensional curves $X(t)$ for $t \in I = [0,T]$ with $T \in \mathbb{R}$ observed on a regular grid. FPCA for such cases has been studied extensively (Dauxois et al., 1982; Silverman, 1996; Besse and Ramsay, 1986; Bosq, 2000) and it was explored for densely observed functional data in Rice and Silverman, 1991; Castro et al., 1986.

Nevertheless, a growing interest is in more complex settings, where images with multidimensional domain are considered. (Wang et al., 2015) describe it as the second wave of FDA, one focusing on voxel and shape analysis. As a result, many different approaches rely on functional regression and FPCA.

With this in mind, two broadly described methodologies are used as a way to approach the high-dimensionality of images. The first methods represent that raw data in terms of a pre-determined system of basis functions, in such cases, the basis functions are carefully chosen as part of the model. The second approach uses the raw data directly and relies on transformations or properties of functions to allow for the estimation of parameters directly.

Image Analysis

An intuitive approach to image analysis in high dimensions is to use a suitable basis expansion which allows for the repurposing of methods originally designed in single dimensions. In this case, the observed raw data $Y_i(s)$ observed on a grid $s \in S \subset \mathbb{R}^3$ is modelled with equation (2.37), and followed by the appropriate use of basis expansion as either a first step in modelling or as an element of functional regression.

For multidimensional functional regression, Wang et al., 2014a use the tensor product of one dimensional Haar wavelet functions to form a 3D basis for the representation of neuroimaging data. Haar wavelets provide a way of overcoming the issue of multicollinearity caused by large spatial correlation among neighbouring voxels whilst modelling sparsity. The choice of basis aimed to identify specific regions relating to an outcome. Their following paper Wang et al., 2017 assumes the image to be piecewise smooth with unknown jumps and edges. The functional regression model was adapted and the image was assumed to be piecewise smooth with unknown jumps and edges, thus relaxing the assumptions of their previous model. Additionally, the previous model was extended for classification.

Park et al., 2016 approach regression in order to identify specific regions by defining a structured way to partition the domain of the regression coefficient. The functional domain selection effectively selects subregions of the brain associated with the outcome. 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.

An alternative approach to basis expansion would be to vectorise the image. Zipunnikov et al., 2011a and Zipunnikov et al., 2011b propose a high-dimensional multilevel FPCA model, aimed for densely-observed images recorded at multiple visits for each subject. The images are vectorised, then the matrix containing subject measurements is partitioned into blocks that then undergo SVD sequentially. 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.

When data is stored as arrays, Li et al., 2019 propose an efficient method for the estimation of FPCA in 3 dimensional images. This method relied on the fact that the inner product of the observations converges to the inner product of the principal component scores. This approach offers an efficient way of estimating the model, especially in high dimensional cases where the direct estimation of the covariance is impossible. An additional benefit is that it does not require the prior basis representation. It is compared against methods vectorising images a priori and has shown to retain more spatial information than those.

Spatio-Temporal Modelling

Originally, the temporal elements of data, whether dense or sparse would be the domain where the principal components are defined. However, as functional data became multivariate, there was an interest in modelling the relationships between functional observations captured at different time intervals. This is commonly called spatio-temporal modelling, and the data could take the forms of curves over time captured at different geographical locations. Previously we have looked at FPCA where the principal components were functions and the scores were random variables. The models described here deal with random functions $X(s,t)$ defined on separate domains $s \in S$ and $t \in T$. In those cases, the principal components were defined over one domain, say S and the scores took the form of a function $\psi(t)$. This class of model can be broadly represented with the equation:

$$X(s,t) = \sum_{j=1}^{\infty} \psi_j(t)\phi_j(s), \tag{2.52}$$

where $\psi_j(t)$ would now be the random score functions varying over time. The definitions of both functions vary depending on the approach proposed.

Greven et al., 2010 proposed a model for the case where $\psi_j(t) = \zeta_{0j} + t\zeta_{1j}$ where ζ_{0j}, ζ_{1j} are random terms. The proposed model is a functional version of a mixed model so a linear structure is imposed on the scores. The method is applied to a brain imaging study designed to analyse differences and changes in brain connectivity in healthy volunteers and multiple sclerosis (MS) patients.

If $Cov\big(\psi_j(t), \psi_j(t')\big) = \lambda_j \rho_j(t - t'; \nu)$, then the model would follow Gromenko et al., 2012 and Gromenko and Kokoszka, 2013 for spatially indexed functional data. Notably, these models were primarily interested in modelling the temporal curvature of the data and capture the spatial correlation in the score functions. In this case $\psi_j(t)$ were the principal components and $\phi_j(s) = \langle X(s,t) - \mu(s,t), \psi_j(t)\rangle$ would have been the scores.

Chen and Müller, 2012 proposed a non-parametric score function that is represented as its on KL decomposition: $\phi_{ij}(t) = \sum_k \eta_{ijk}\zeta_{ijk}(t)$ with orthogonal basis functions $\zeta_{ijk}(t)$ and the corresponding coefficients η_{ijk}. $X(s,t)$ is represented as as a product of these score functions and a time-varying orthogonal basis function $\phi(s|t)$ serves as the principal components.

Park and Staicu, 2015 present a flexible model with non-parametric score functions as in Chen and Müller, 2012 with the main difference in how they define the orthogonal basis to act as principal components. They define it to be the eigenvectors of the marginal covariance function of the random process together with a residual process.

Chapter 3

Neuroimaging Data Background

In this thesis, we model high dimensional data in the form of brain magnetic resonance (MR) imaging. This chapter will introduce MR imaging by describing what it is, how it is acquired and preprocessed. We will then discuss the two datasets used in the thesis alongside previous analysis and ongoing research questions.

3.1 Introduction to Magnetic Resonance Imaging

Neuroimaging includes the use of various techniques to produce the image of the structure, function or pharmacology of the brain. These techniques include x-ray computed tomography, magnetic resonance imaging (MRI) or position emission tomography (PET), each specialised for a different purpose, with some being less invasive than others. In our case, we are particularly interested in MRI, with the following introduction supported by Prince and Links, 2015 and Poldrack et al., 2011.

The two modes MR imaging that are most pertinent to the thesis are structural imaging and functional imaging. The former applies a host of pulse sequences to the brain that allows for the observation of its structure facilitating tasks such as diagnosis of intracranial disease. The latter measures an aspect of brain function, often with a view to understanding the relationship between activity in certain brain areas and specific mental functions. It uses oxygenation-sensitive pulse sequences to image blood oxygenation in the brain with high oxygenation correlating to brain activity.

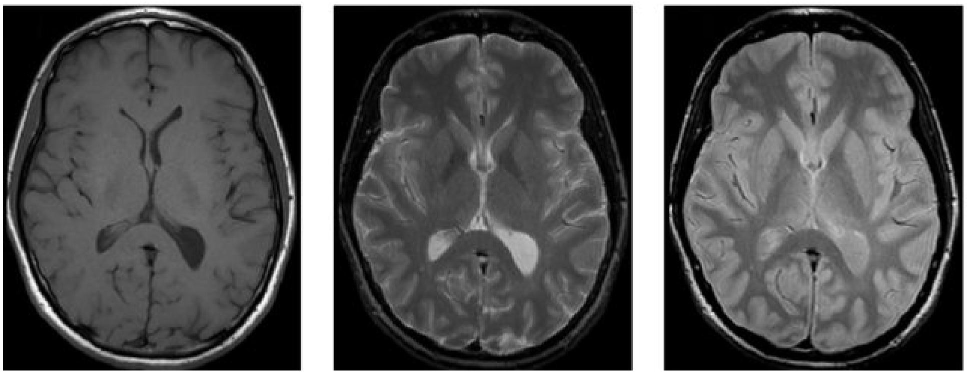


Figure 3.1: Examples of structural MRI: (a)T_1-weighted, (b) T_2-weighted and (c) P_D-weighted. Elnakib, 2013

The parts of the brain that are often observed within structural imaging, are white matter, grey matter and cerebral spinal fluid (CSF). White matter consists of the long axons of neurons that conduct electrical signals to more distant regions of the brain and spinal cord. Grey matter consist of neuronal cell bodies and their dendrites, which are short protrusions communicating with neurons close by. In MRI, one can create different images to highlight each tissue type by exploiting its NMR properties. Structural imaging can thus be further subdivided into T_1-weighted, T_2-weighted and P_D-weighted (P_D standing for proton density). These can be seen in Figures (3.1) and (3.2). T_1-weighted MR images provide a clear view of brain anatomy and structure, making them useful in analysing soft tissue and identifying damage. T_2-weighted images are used to measure white matter and cerebrospinal fluid in the brain, thus they are more suited to measure fluid rather than soft tissues.

Functional MRI can be further subdivided into two categories, task and resting state fMRI. Task fMRI has patients perform tasks within an MRI scanner to learn how the brain responds to various stimuli. Resting-state fMRI, as the name suggests, allows for the patients to rest within the machine and instead of a response to stimulus, it allows to look for spontaneous activity in the brain, are there any parts of the brain that are connected. In both cases, however, the fMRI measures signals from the changes in oxygenation that are referred to as the blood oxygenation level dependent (BOLD) signals.

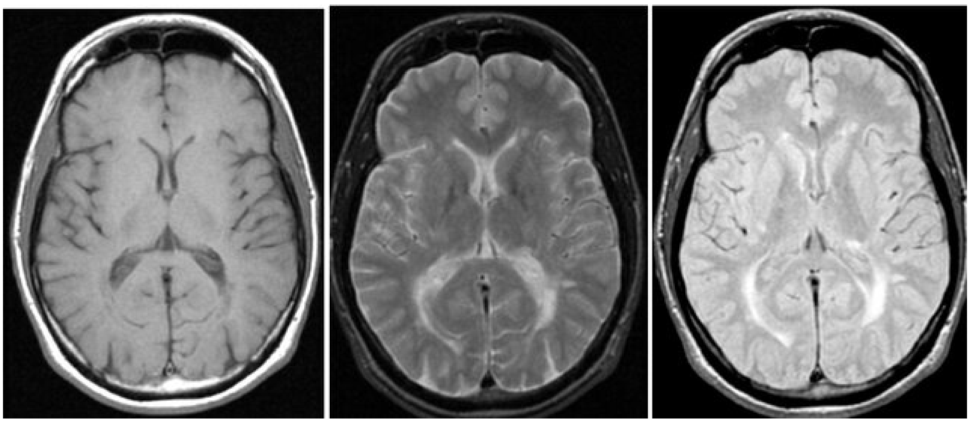


Figure 3.2: Examples of structural MRI: (a)T_1-weighted, (b) T_2-weighted and (c) P_D-weighted. Ashton and Du, 2004

Challenges in MR Image Analysis

The analysis of an MR image is subject to a number of factors. Firstly, on an individual level, the data is liable to a number of artifacts, such as those caused by head motion or fluctuations in signal sensing. Secondly, on a sample level, there are many sources of variability in the data, including variability between individuals and time within individuals. Thirdly, the data is of rather high dimension. A T_1-weighted MRI can be $250 \times 250 \times 250$ voxels large, multiplied by the number of longitudinal observations and the number of subjects can lead for the complete dataset to be tens if not hundreds of gigabytes.

Individual and sample level challenges most commonly include:

- **Artifacts**: Most common form of image distortion is geometric warping or complete loss of signal (most commonly in fMRI), this arises when the gradient strength is not uniform across the entire field of view.
- **Signal-to-noise ratio**: noise arises from statistical fluctuation of the signal sensed by the receiver coils artifacts. This is present in all types of MR images.
- **Subject variability**: as each patient has unique anatomy, a voxel coordinate may not correspond to anatomical features of the brain across multiple subjects images.
- **Motion**: MR acquisition time can take from 10 minutes up to an hour depending on the specific type of imaging done. In both types of imaging, however, motion of the head can reduce the quality of resulting images.

Examples of movement related artifacts and signal-to-noise ratio can be seen in Figures (3.3) and (3.4), respectively.

Finally, when analysing an image, one should be aware that voxels do not indicate the tissue type directly, but they indicate tissue types relative to each other given the contrasting mechanism used (T_1, T_2 or P_D). Voxel intensity often does refer to specific tissue composition, however it does not measure brain matter volume directly.

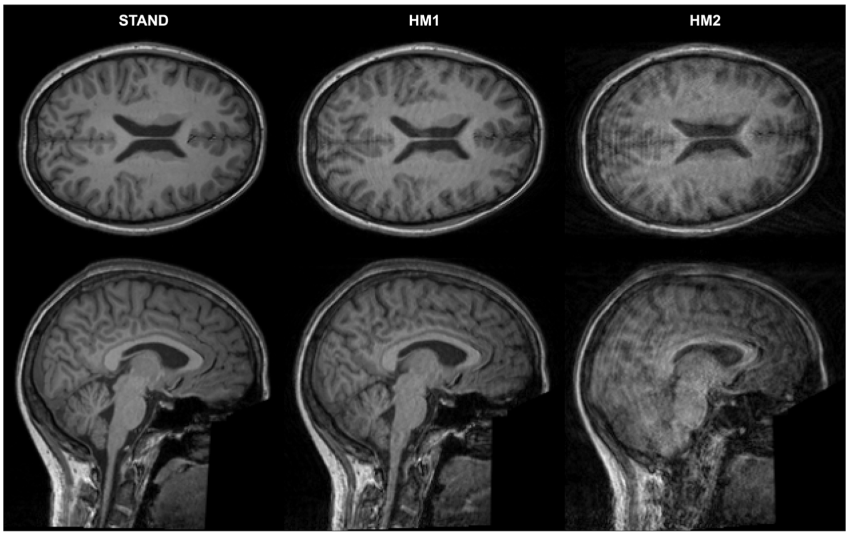


Figure 3.3: Clean and motion-corrupted images of one representative participant. One axial and one sagittal slice are presented for the standard (STAND) scan, and for scans with low (HM1) and high levels of head motion (HM2). For this participant, the STAND scan was labelled as good (score 1), the HM1 scan as medium (score 2), and the HM2 scan as bad (score 3) quality image from the point of view of clinical diagnostic use. Narai et al., 2022

3.2 MRI Preprocessing Steps

Prior to analysis, images must go through a number of steps to either remove existing image distortions, insignificant tissue and to account for variability in subject anatomy. Many different preprocessing pipelines have been developed to suit multiple analysis goals, and this step should be considered when interpreting results.

Whilst there are many software packages that can be used for brain MRI preprocessing, the ones used in this thesis are FSL (Smith et al., 2004; Jenkinson et al., 2012) and ANTs (Avants et al., 2009). Each software contains a collection of relevant tools that are frequently used in a pipeline. The following section will go over the most pertinent elements to the thesis.

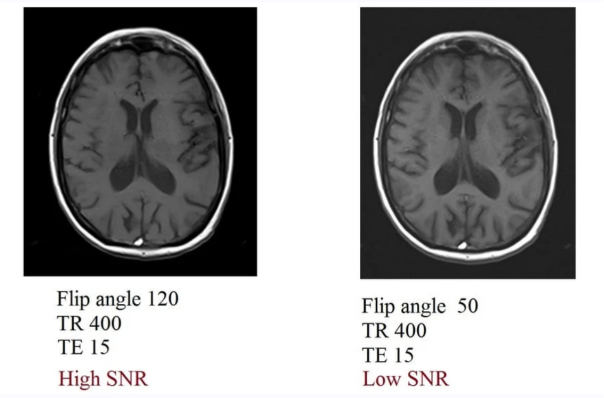


Figure 3.4: Examples of high and low signal-to-noise ratio in T_1 MR images.

Bias Correction

Bias field signal is a low-frequency and smooth signal that corrupts MRI images and is particularly prevalent in ones produced by old MRI machines (Juntu et al., 2008). An example of a bias field can be found in Figure (3.5). It can be a potential confounder in analysis tasks that depend on voxel values. A popular bias field correction method was introduced initially Sled et al., 1998 and is based on fitting b-splines to represent the bias field. It was later improved upon by Tustison et al., 2014 and this is included in the current version of ANTs and was used in this thesis.

Registration

Brain registration is the act of aligning images across subjects using a variety of transformations such that for each image, a voxels location corresponds to the same anatomical location between patients. Registration accounts for different rotations of the brain and the variability of patient brains.

The transformations that make up registration can be simple rigid body and affine transformations such as skews and shears along different axies. Non-linear transformations can be subject to constraints such as basis functions, regularization and topology-preservation (Ashburner and Friston, 2000).

Brain volume and other anatomical features heavily depend on age, gender and geographic region and thus vary across populations. This heterogeneity could limit the generalizability of many studies and is mitigated by mapping the images onto a common template.

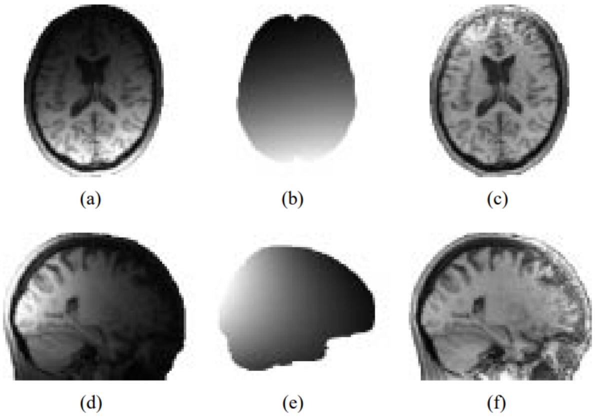


Figure 3.5: Intensity nonuniformity correction of a surface coil MR scan: (a) and (d) transaxial and sagittal views of uncorrected data; (b) and (e) nonuniformity field estimated by the N3 method; (c) and (f) corrected data. Image from Sled et al., 1998.

These templates are often derived from a set of images obtained from a large study. Fonov et al., 2011 introduced such a template from an atlas of images that averaged (over the population) the intensity, average shape, left-right symmetry, high level of anatomical detail and compatibility with previous atlases (Evans, 2006; Almli et al., 2007; Lancaster et al., 2007; Mazziotta et al., 1995). This atlas is referred to as MNI152 and was used as part of the ANTs software.

Image Normalization

Image or intensity normalization aims to standardise the relative contrast of the observed pixels. It is an important step as many analysis methods make strong assumptions about the underlying intensity ranges within an image what tissue they might correspond to. Additionally, many methods assume the data to have been sampled independently and identically distributed from a fixed distribution which is not always the case. These variations can stem from differences in the protocols of various MRI scan acquisitions, the different manufacturers and scanner-models, and also due to subject disease state. One subjects image at different time points can have different tissue intensity.

Normalization addresses this problem by mapping the raw image intensity values into a standardised range. This transformation results in a standard scale where intensities in the transformed images have consistent tissue meanings and standard window settings can be determined for different tissues. An example of normalized intensities can be seen

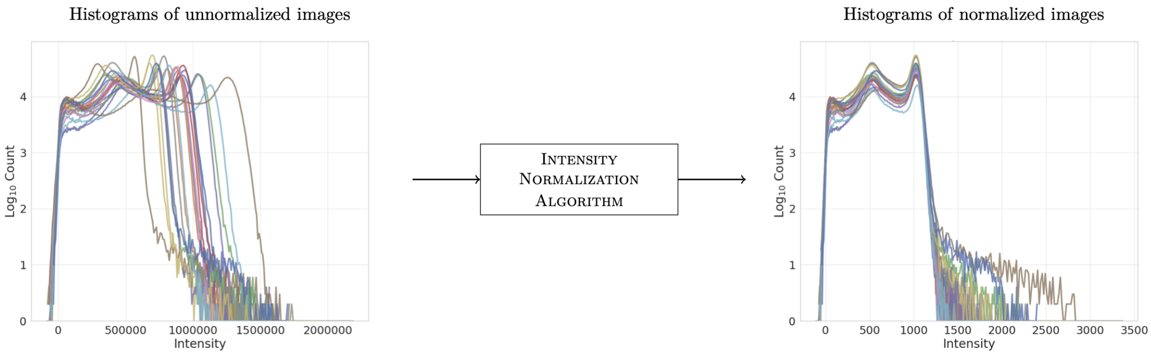


Figure 3.6: An example of image intensities before and after normalization. Image sourced from Reinhold et al., 2019

in Figure (3.6).

Different methods can be used to achieve a result and a useful tool in this thesis was the repository of methods reviewed by Reinhold et al., 2019. The specific method used was the piecewise linear histogram matching introduced and evaluated in Nyul et al., 2000; Shah et al., 2011. This method addresses the normalization problem by learning a standard histogram for a set of contrast images and linearly mapping the intensities of each image to this standard histogram. It is particularly useful as it normalizes the whole sample of images relative to each other.

Brain Extraction

Brain extraction is, as the name suggests, the process of differentiating the brain tissue from non-brain tissue. It is most frequently used on high resolution magnetic resonance (MR) images as they often depict non-brain matter such as eyeballs, bone and muscle. In contrast, functional images, because of their rapid acquisition rarely depict non-brain tissue. Where it is appropriate, then, removing non-brain matter will allow for the analysis to focus on regions of interest.

There are many ways of approaching this problem. Lemieux et al., 1999 suggests a series of thresh-holding and morphology steps, with each step carefully tuned to overcome specific problems, such as the thin strands joining brain to non-brain after thresholding. Whilst very accurate, this method proved limiting due to its narrow range of applications. Dale et al., 1999 suggests fitting the image to a surface model composed of a triangular mesh. Many other approaches were and continue to be developed (Stella Atkins and

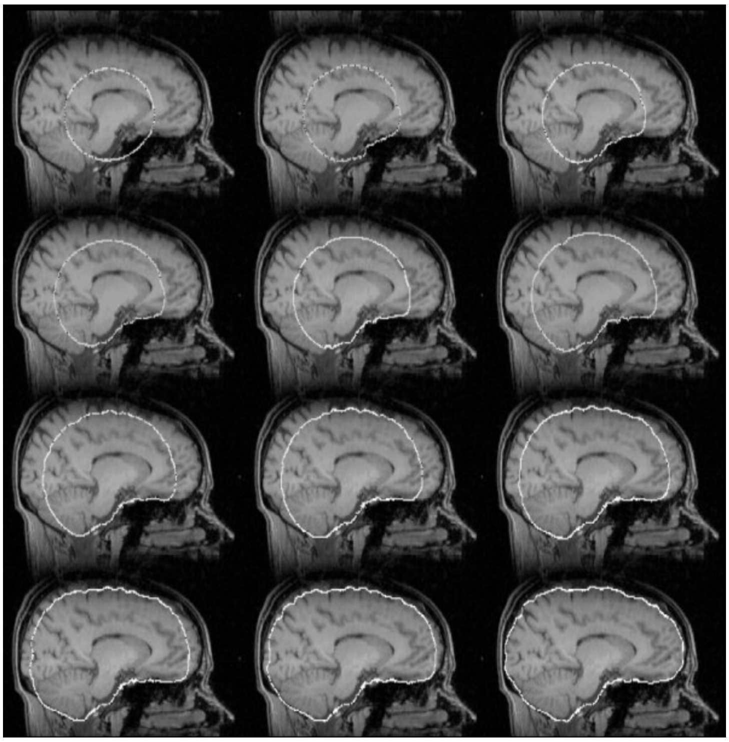


Figure 3.7: Example of iterative surface model development. The dark points within the model outline are vertices. Image from Smith, 2002

Mackiewich, 1998; Wang et al., 2014b; Carass et al., 2011; Shattuck et al., 2001).

One method, which is frequently used – introduced by Smith, 2002 – uses a deformable model that evolves a surface to fit the brain boundary by accounting for surface smoothness and voxel intensity changes in the surface vicinity. An example of this surface developing can be seen in Figure (3.7) This approach is used by the FSL software as the Brain Extraction Tool and remains popular partially due to its low requirement for human monitoring. In practice, one needs to choose a fractional intensity threshold in the range [0,1] where smaller values give larger brain outline estimates.

Smoothing

Some images may be smoothed prior to analysis. This is mostly only used on functional images to suppress spatially random noise and enhance the signal-to-noise ratio. Functional MR analysis has special constraints due to the spatially varying nature over voxels that can cause intrinsic autocorrelations. Friston et al., 2000 discusses how smoothing ensures that the bias is small whilst maintaining a reasonable degree of efficiency.

Smoothing, then, is a common step in fMRI preprocessing and a Gaussian kernel, whose size and intensity is determined relative to the data, is often used. Smoothing is not the only step that might be specific to fMRI as head motion may have a large effect on the analysis of the series of images (van Dijk et al., 2012). Kassinopoulos and Mitsis, 2022 provide additional further details of evaluating suitable methods for fMRI preprocessing.

3.3 Decision fMRI

The following section will discuss the first of two datasets that are analysed in this thesis. This task fMRI dataset, first published by Mohr et al., 2010b, was used to learn which regions of the brain are active in risk-averse individuals.

3.3.1 Study Design

The data is sourced from a previous study (Mohr et al., 2010b) investigating the mechanisms behind decision making and risk. The authors designed a investment decision task that uses streams of (past) returns as stimuli to understand human responses to different levels of risk, and to use their responses to investigate the two competing models for risk in neuroscience. The experiment involved a cohort of 22 young subjects (age 18–35 years, equal gender distribution, native German speakers, right-handed and had no history of neurological or psychiatric disease). Three participants were excluded initially for excessive head motion and a further two subjects were excluded for corrupted image files.

Each trial task consisted of two phases: the presentation of a return stream followed by a decision or a subjective judgment task (Fig. 3.8). In the first phase, the return stream consisted of 10 returns on a previous hypothetical investment giving information about its past performance. Each individual return was presented for 2 seconds and the total return stream takes 20 seconds to complete. During the experiment, each return stream was independent of the others and described as the past performance of a new investment. The streams were drawn from Gaussian distribution with varying means (6 %, 9 %, and 12 %) and standard deviations (1 %, 5 %, and 9 %), resulting in nine different combinations of means and standard deviations.

In the second phase, given the return stream, subjects were asked to do one of the three

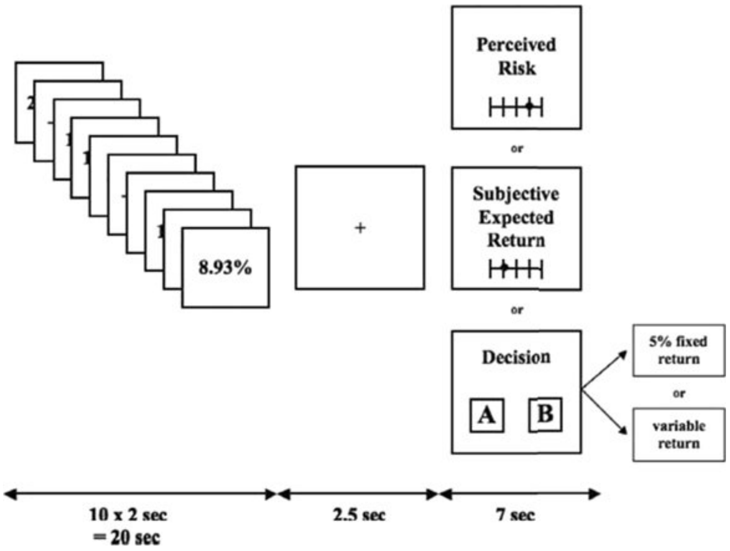


Figure 3.8: Risk Perception and Investment Decision Task Mohr et al., 2010a.

tasks, these were chosen to be able to investigate choice, perceived risk and subjective expected return. They are:

1. Decision task: the subject has to choose between an investment with a 5% fixed return (safe investment) and an investment represented by the return stream they just saw (risky investment).
2. Expected return: The subjects would choose the subjective expected return to be an integer value between -5% and +15%
3. Perceived risk: Subjects would give their perceived risk on a scale from 0 (no risk) to 100 (maximum risk).

This was done without knowing which of the three tasks they would be performing before the stream began. Subjects performed each task (decision, subjective expected return and perceived risk) 27 times giving a total of 81 trials. Figure (3.8) summarises the study design.

During the experiment, fMRI data were acquired on a 1.5 T Magnetom Sonata fMRI system equipped with a standard head coil. A vacuum pad was used to minimize head motion. Functional images were acquired using a BOLD-sensitive T2-weighted echo-planar imaging (EPI) sequence. This resulted in 1400 observations of a 3-dimensional (91 × 109 × 91) array that represents the Blood Oxygenation Level Dependent (BOLD) signals. The data was initially pre-processed with FSL 4.0, which included motion correction and slice-time correction. Additionally, images were normalized into a standard space.

Below are the results for each of the tasks completed by the subjects. We want to use this to help us in the analysis of the data. Table (3.1) shows the prevalence of risky choices in all types of decision tasks. The first row will do this for expected return judgment and the second row will do this for perceived risk judgment. Subject risk prevalence in task types 3 and 4 is plotted using box-and-whisker plots and can be found in the appendix as Table (A.1) and Table (A.2).

Task Type 1														
Index	3	4	13	20	24	36	39	43	49	58	59	72	76	77
Risk %	0.8	0.6	0.53	0.13	1	0.53	0.93	0.2	0.86	0.8	0.86	0.26	0.86	0.13
Task Type 2														
Index	5	15	18	23	32	34	37	52	54	62	64	71	79	
Risk %	0.8	0.2	0.93	0.93	0.6	1	0.53	0.93	0.13	0.53	0.13	0.93	0.8	

Table 3.1: The prevalence of risky labels in the two task types by task index.

3.3.2 Results from the Study

The subjects' statements for perceived risk and subjective expected return were used to identify which mathematical model best translates the 10 presented returns into predictions for perceived risk and subjective expected return on an individual level. Thus, one can use these models to predict perceived risk and subjective expected return during the choice between the risky and the safe investment, where they are otherwise unobservable.

The data consists of brain fMRI images of 15 subjects, taken whilst they performed 81 financial decision-making tasks or experiments. The images were acquired every 2.5 s during the investigation, providing each subject with a time series of 1360 three-dimensional images representing their brain activity, an example of one subject's image can be seen in Figure (4.5). The images represent the BOLD signals. Each patient also received a clinically-derived risk score at the end. In this study, all participants were risk averse.

Mohr et al., 2010a fitted the risk–return model with expected return and standard deviation in the same way we fitted the psychological risk–return model and also tested how well the resulting risk weights could explain the choices if we assume a deterministic decision rule. It was able to predict on average 85% of the choices and found a significant correlation between perceived risk and the BOLD response during decisions in right anterior insula (aINS) and right OFC

3.3.3 Previous Analysis

This dataset was first published by Mohr et al., 2010b and subsequently analysed using various FDA methods by van Bömmel et al., 2013, Chen et al., 2015 and finally by Li et al., 2019. The aim of the aforementioned papers was to recover active brain areas that are associated with risk assessment and to use elements of the model to predict on subjects risk attitude parameter. Previous studies found correlations between risk attitude and risk-related brain activity in the lateral orbitofrontal cortex (lOFC) for risk-averse individuals and in the medial orbitofrontal cortex (mOFC) for risk-seeking individuals (Tobler et al., 2007). There, risk-averse individuals would weight the risk associated with an investment to discount it's overall value whereas risk-seeking individuals would have a as negligible risk weight, meaning their perceived value of an investment was not altered significantly by a change in risk. Another study (Mohr et al., 2010b) found that inter-individual differences in decision-related brain activity in the lOFC and correlated it with inter-individual differences in risk attitudes independent of the current level of risk. The authors showed that the value signal in the ventrolateral prefrontal cortex (VLPFC) increased with risk in risk-seeking individuals and decreased with risk in risk-averse individuals, thereby reflecting the risk attitude. Additionally, Chen et al., 2015 have found the dorsolateral prefrontal cortex (DLPFC) and the anterior insula (aINS) to

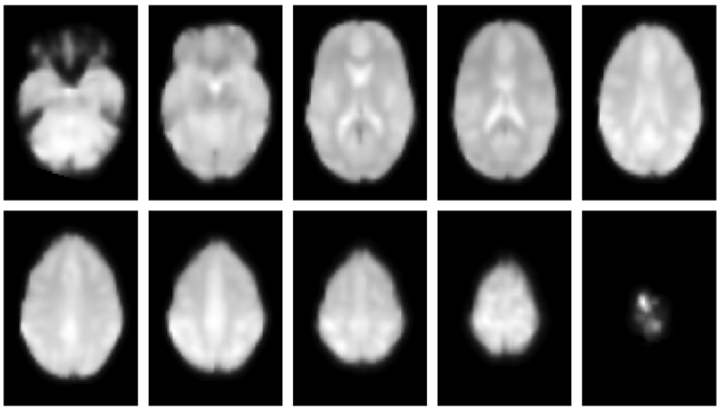


Figure 3.9: Raw data slices for one patient. This represents the data $Y_{1,(s_1,s_2,s_3),t}$ where $t = 1$, $s_1, s_2 \in [0, 91]$ and $s_3 \in \{25, 30, 40, 45, 50, 55, 60, 65, 70, 80)\}$.

be correlated with decision making and risk.

van Bömmel et al., 2013 hypothesize that the temporal variability of components corresponding to factors in brain regions related to value processing is correlated with the risk attitude of individuals. Based on this, we are interested if we can identify regions of interest (ROIs) and relate their temporal activity to the clinical risk score given to each patient. Given the fact that each patient completed multiple tasks, we would like to find out if individual trends in behaviour would impact the assumption posed by Li et al., 2019 stating that each task could be treated as an individual observation. Finally, we are interested in developing a model that maintains the spatial relationships within each image and can simultaneously represent the time series.

3.4 Alzheimer's Disease Neuroimaging Initiative

The second dataset used in this project is a set of images published by the Alzheimer's Disease Neuroimaging Initiative (ADNI)(Petersen et al., 2010) which follows multiple patient cohorts over the course of years to study biomarkers related to Alzheimer's Dementia (AD).

AD is a chronic neurodegenerative disorder with progressive impairment of the memory and other important mental functions. The condition is characterized by morphological and molecular changes of the brain, ultimately leading to cognitive and behavioral decline. As age is a major risk factor of the condition, it was increasingly important to study its onset and progression due to an increase in life expectancy and an ageing population.

The first major study to do this over multiple sites was ADNI (Mueller et al., 2005). It collected the genetic, neuroimaging and biochemical biomarkers on an elderly population with the aim of improving diagnostic criteria and establishing relationships between disease onset and progression between a variety of biomarkers to better understand potential treatments. The first cohort results were published by Petersen et al., 2010. The dataset contains clinical, neuroimaging, and cognitive data, as well as biofluid samples. Since the publication of the ADNI 1, several subsequent studies have been published: ADNI GO (2009-2011), ADNI 2 (2011-2017) and ADNI 3 (2017-2022). Currently, there is a

Month	**0**	**6**	**12**	**18**	**24**	**36**
CN	134	137	138	6	136	132
MCI	148	139	123	102	92	78
AD	97	106	121	40	153	72
Total	379	382	382	148	382	282
na	3	0	0	234	0	100

Table 3.2: Number of patient with each diagnosis at each visit time.

transition to a new study ADNI 4 (Weiner et al., 2023).

ADNI had a large impact on the establishment of multi-center, large scale trials (Weiner and ADNI, 2013; Weiner et al., 2017) and has been used widely for method development. Most recently, a review of publications between 2021 and 2022 Veitch et al., 2024 has identified 1459 publications in that year using the dataset. This is part due to the many collections of datasets that can be readily downloaded. The dataset of particular interest to this thesis is the standardised 1.5T MRIs set (Wyman et al., 2013) following 382 subjects over 3 years with visits scheduled a 6 month intervals. For each visit, subject data is available regarding their diagnosis, demographics, genetics and other. The number of patients with each diagnosis sub-type (CN, MCI and AD) are summarised in Table (3.2). Subjects' age at screening divided by diagnosis is plotted in Figure (3.10).

3.4.1 Neuroimaging and Alzheimer's Dementia

Structural MR imaging is used in the diagnosis process along with the monitoring of disease progression amongst patients. The onset of Alzheimer's dementia has been associated with accelerated atrophy is several brain regions, particularly in the medial temporal lobe with concurrent expansion of the ventricles (Park and Reuter-Lorenz, 2009; Jack et al., 1992; Fox et al., 1996). A volume reduction in the hippocampus, a sub-region of the temporal lobe has been associated with dementia (Thompson et al., 2004). As such, volume reduction in particular brain regions can be considered as an imaging biomarker used to investigate the rate of brain deterioration. This has been quantified with different techniques, one counting the neuronal cell loss (West et al., 1994) or by computing the brain volume loss directly (Leong et al., 2017).

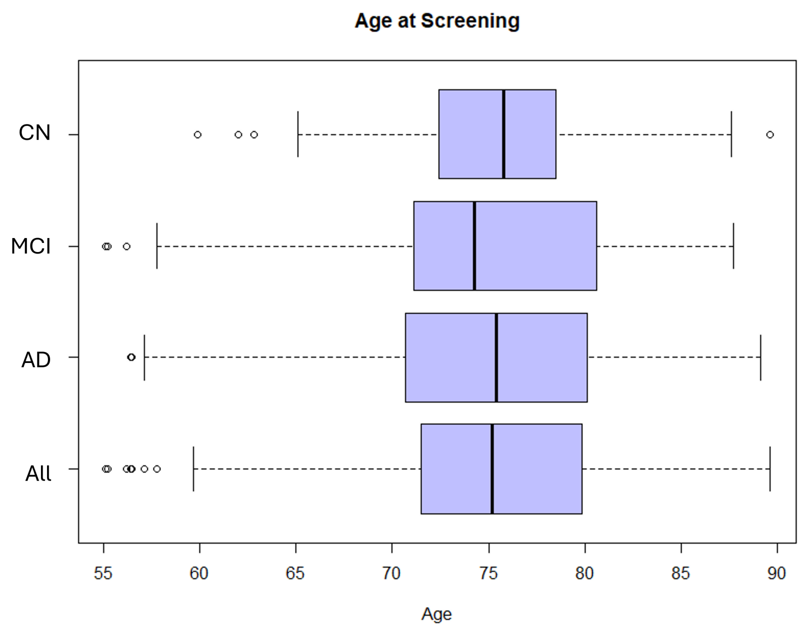


Figure 3.10: Boxplots of ADNI subjects' age at screening by diagnosis.

3.4.2 Analysis of ADNI Images

Given the importance of neuroimaging in AD diagnosis and monitoring, several initiatives have sought to collect large datasets of images and other relevant clinical measurements to obtain more insights about the disease progression. Such data has allowed us to model the trajectory of the disease over time, helping develop methods for precise and early diagnosis. We are interested in methods that fall into the categories of classification of and prediction of diagnosis or future decline. We will consider both statistical and machine learning (ML) approaches that work either on full images or on biomarkers extracted from them.

Modelling of Longitudinal Data

MRI biomarkers are frequently modelled with mixed models to find associations between patient profiles and disease outcomes (Chen et al., 2021). However linear mixed models can have some limitations: (1) the parametric models can be limited when modelling complex nonlinear trends of longitudinal data; (2) missing points can make the model difficult to estimate or completely unidentifiable. Non-linearity is important with regards to AD as many biomarkers have shown to have complex trajectories dependent on age and genetic status (Jack et al., 2012). Furthermore as rate of decline is non-linear, using the diagnosis time or time of enrollment may not directly reflect the patients trajectory (Milliken and Edland, 2000).

Nonlinearity has been addressed by several approaches. Some models include piecewise models (Gerstorf et al., 2010), mixed effect change point models (Hall et al., 2000) or a flexible sigmoidal model (Capuano et al., 2018). FDA provides a non-parametric solution to the problem of modelling longitudinal data. One can use FPCA to model longitudinal trajectories and relate them to subject specific scores (Shi et al., 2021; Yao et al., 2005). Functional mixed models (Guo, 2002) have been extended to account for random variables over different domains (Happ and Greven, 2018) or to jointly model multivariate functional data with survival outcomes (Li et al., 2022; Zou et al., 2023). The point of sparsity has been addressed using a Bayesian approach, which has been applied both in standard mixed models (Li et al., 2018) and functional approaches (Yao et al., 2005; Thompson and Rosen, 2008).

Image Analysis

Predicting or classifying disease status from imaging is well established in the context of Alzheimer's Dementia. Imaging biomarkers have been widely used in predicting the onset of dementia and many methods have been considered for this task, often extracted features are used in a classification model to predict the presence of AD. This has been achieved by both statistical and machine learning approaches.

Mofrad et al., 2021 provide a framework for the extraction of brain biomarkers and applying them in a mixed model for the prediction of patient diagnosis at a given time. A similar approach for the prediction of time to conversion was presented in Guerrero et al., 2016. At the same time, extracted biomarkers can be used in non linear ML methods. Review articles on this topic in the context of ADNI have been recently published by Ansart et al., 2021; Fouladi et al., 2022; Rowe et al., 2021; Grueso and Viejo-Sobera, 2021. These publications highlight that support vector machines (SVMs) and neural networks are the most common approaches. Methods using a complete image without preprocessing are more recent, and mainly include NNs, specifically, convolutional neural networks (CNNs) (Fouladi et al., 2022).

Functional approaches could be broadly categorized into FPCA and functional regression. Palma et al., 2020 uses a b-spline representation as a summary of the 3D image to then estimate a quantile regression model for brain age. Another approach was functional logistic regression Wang et al., 2017 on full FDG PET scans where the fitted coefficient was composed of wavelet functions.

3.4.3 Data Structure and Preprocessing

The data was downloaded in the NIFTI format, which is a typical data structure for MRI alongside DICOM. The data is then pre-processed using the clinica pipeline to first order it in the BIDS format and then preprocess the images themselves. The image preprocessing was done using the clinica t1-processing (Routier et al., 2021; Wen et al., 2020). More precisely, bias field correction was applied using the N4ITK method (Tustison et al., 2010). The images were aligned to a common imaging space that aligns each image such that the anatomical regions of the brain corresponded to the same area in the image space. This is done using affine registration via the SyN algorithm (Avants et al., 2008) from the ANTs software (Avants et al., 2014) and the common imaging space is the MNI space with the ICBM 2009c nonlinear symmetric template (Fonov et al., 2011). The registered images were further cropped to remove the background resulting in images of size 169×208×179, with 1 mm isotropic voxels.

Chapter 4

Spatio-Temporal Functional Principal Component Analysis

4.1 Introduction

In this chapter, we are interested in an fMRI dataset collected on a dense time grid introduced in Chapter 3. We propose a functional model using the FDA framework for the dimensionality reduction of images observed over time. The model builds on previous work of Li et al., 2019, that had introduced an efficient estimation method for the estimation of a standard FPCA model on images stored as arrays. It takes inspiration from Park and Staicu, 2015 by using a double Karhunen-Loève expansion to achieve a decomposition of the image into two functions: a time-invariant PC function and a random, time varying score that can be used in further analysis. This work is motivated by task fMRI data introduced in Chapter 3, where 15 healthy subjects performed financial decision-making tasks and the response of interest was a subject's financial risk propensity. The model can be used to obtain subject specific score functions over time which can be used further to relate active brain regions to a subject's risk preference.

Task fMRI studies aim to find relations between activity in a particular brain region and a subject response. Most commonly, such datasets are analysed with a per-voxel approach which includes generalized linear models (GLMs), independent component analysis (ICA) and time series models (Friston et al., 1995 and Monti, 2011). These methods tend to focus on the temporal correlation treating each voxel's signal as spatially independent from the others. From the very first application of ICA to fMRI there has been a discussion on modelling spatial or temporal dependence (Mckeown et al., 1998; Petersen et al., 2000). On one hand, ICA provides a model for brain activity over time at particular voxels and on the other, different dimensionality reduction methods, can model the spatial dependence.

FDA methods are particularly useful as treating the image as a functional object inherently models the spatial correlation of adjacent pixels. As such, modelling neuroimages using FDA has grown in popularity (Wang et al., 2015). Many of these methods rely on dimensionality reduction, such as FPCA, where the PCs form spatial latent variables that represent some element of the variation in pixels and the random scores can be used in further analysis such as association or prediction. In the case of fMRI, early implementation of functional methods was presented by Zipunnikov et al., 2011a, where the image was vectorised to estimate a 2 dimensional covariance matrix. Another approach in Chen et al., 2015 first represented the data using b-spline vectors which is a common method in FDA also described in Chapter 8 of Ramsay and Silverman, 2005. Most recently, Li et al., 2019 presented a new estimation method for FPCA on images stored as arrays. However, these methods do not consider the temporal element of fMRI data and either consider a single time point, as is the case for Zipunnikov et al., 2011a; Chen et al., 2015 or they combine multiple time points into one image by taking their difference as done in Li et al., 2019.

Methods that do consider spatial dependence are often then limited by not considering time. Few methods consider both. One approach by van Bömmel et al., 2013 uses methods introduced in Park et al., 2009 on patches of images, where the patch is decomposed into a spatial component and an autoregressive model is fitted to the scores. However this approach relies on parametric assumptions and is computationally expensive. Other methods have been developed on low dimensional data and are not adapted to images. Two non-parametric approaches that allow for flexible modelling include Chen and Müller, 2012; Park and Staicu, 2015, however both are limited to 1-dimensional curves over time and cannot be directly implemented to imaging data.

In this chapter, we propose a functional model using the FDA framework for the dimensionality reduction of images over time, with a dense temporal grid. The time-series of images is decomposed into a linear product of time invariant principal components and subject specific score functions that are allowed to vary over time. Our model is computationally efficient, as we propose an estimation method based on the algorithm introduced in Li et al., 2019 that circumvents calculating the 8-dimensional covariance matrix. Our novel contributions are as follows: firstly, we propose a model for dimensionality reduction of high-dimensional datasets captured in space and time. Secondly, we implement a computationally efficient estimation algorithm using a singular value decomposition and study its performance via simulation. The data analysis compares our approach with Li et al., 2019 and recovers active brain regions and associate their activity over time with subject's risk attitude.

4.2 Methods

Define a random function $X(s,t)$ where $s \in S = [0,S_1] \times [0,S_2] \times [0,S_3]$ forms a bounded 3-dimensional space, $t \in \mathcal{T} = [0,T]$ and $S \cup \mathcal{T} \subseteq \mathbb{R}^4$. The function $X(s,t)$ lies in $L^2(S \cup \mathcal{T})$. As $X(s,t)$ is integrable, there exists a unique function $\mu \in L^2$ such that $\mu(s,t) = \mathbb{E}[X(s,t)]$. Hence, the function $X(s,t)$ can be decomposed into the mean $\mu(s,t)$ and some function $U(s,t) \in L^2(S \cup \mathcal{T})$ such that $X(s,t) = \mu(s,t) + U(s,t)$. If we were to model this function using standard FPCA described in Chapter 2, $X(s,t)$ would be expressed as:

$$X(s,t) = \mu(s,t) + \sum_{l=1}^{\infty} \psi_l \phi_l(s,t) \tag{4.1}$$

where $\phi_l(s,t)$ are the PCs and ψ_l would be the scores. However, we are interested a representation where $X(s,t)$ is a product of two functions, a time-invariant PC function defined over s and a random score function over t. Existing models introduced by Park and Staicu, 2015 cannot be estimated in high dimensions, due to lack of tools for eigen-decomposition of 6 dimensional arrays, and are investigated in detail in Chapter 5. In this section, we define a model for such a decomposition that can be estimated in high dimensions.

Consider the marginal function $U(s) = \int U(s,t)g(t)dt$ where $g(t)$ is the sampling density function over t, $g(t)$ is continuous and $sup_{t \in \mathcal{T}}(g(t)) < \infty$. The function $U(s)$ has a covariance $\nu(s,s') = \mathbb{E}\Big[U(s)U(s')\Big]$ which is the kernel to the covariance operator:

$$\varsigma(f)(s') = \mathbb{E}\Big[\langle U(s), f(s)\rangle U(s')\Big], \tag{4.2}$$

for any functions $f(s), g(s) \in L^2(S)$. The covariance operator is symmetric as $\nu(s,s') = \nu(s',s)$ and positive semi definite, thus has strictly positive eigenvalues λ_l with $\sum_{l=1}^{\infty} \lambda_l < \infty$. From the eigendecomposition problem:

$$\varsigma(\phi_l)(s') = \int_S \nu\big(s,s'\big)\phi_l(s)ds = \lambda_l \phi_l(s'), \tag{4.3}$$

the kernel of the covariance operator, ν can be written as:

$$\nu(s,s') = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(s'),$$

with $\phi_j(s)$ denoting the eigenfunctions. The eigenfunctions form a time-invariant orthonormal basis in $L^2(S)$ and optimise the minimisation of:

$$\begin{aligned} MSE(\theta_1(\cdot),\dots,\theta_K(\cdot)) &= \mathbb{E}\Big|\Big|\int U(\cdot,t)g(t)dt - \sum_{k=1}^{K}\langle\int U(\cdot,t)g(t)dt, \theta_k(\cdot)\rangle\theta_k(\cdot)\Big|\Big|^2. \\ &= \mathbb{E}\Big|\Big|\int U(\cdot,t)g(t)dt - \sum_{k=1}^{K}\int\Big(\int U(\cdot,t)\theta_k(\cdot)d\cdot\Big)g(t)\theta_k(\cdot)dt\Big|\Big|^2 \\ &= \mathbb{E}\Big|\Big|\int\Bigg(U(\cdot,t) - \sum_{k=1}^{K}\Big(\int U(\cdot,t)\theta_k(\cdot)d\cdot\Big)\theta_k(\cdot)\Bigg)g(t)dt\Big|\Big|^2 \\ &= \int \mathbb{E}\Big|\Big|U(\cdot,t) - \sum_{k=1}^{K}\Big(\int U(\cdot,t)\theta_k(\cdot)d\cdot\Big)\theta_k(\cdot)\Big|\Big|^2 g^2(t)dt \\ &\text{since g(t) is deterministic} \\ &= \int \mathbb{E}\Big|\Big|U_i(\cdot,t) - \sum_{k=1}^{K}\langle U_i(\cdot,t), \theta_k(\cdot)\rangle\theta_k(\cdot)\Big|\Big|^2 g^2(t)dt. \end{aligned}$$

By applying Mercer's theorem and the Karhunen-Loève theorem, the process $U(s)$ can be expressed as an infinite linear combination of the deterministic eigenfunctions $\phi_l(s)$ of

$\varsigma(U)(s)$ with random uncorrelated weights $\omega_l = \langle U, \phi_l \rangle$:

$$U(s) = \sum_{l=1}^{\infty} \omega_l \phi_l(s). \tag{4.4}$$

We propose this new $\phi_l(s)$ to be the basis function of new decomposition of $U(s,t)$ together with the random score functions $\psi_l(t) = \langle U(s,t), \phi_l(s) \rangle$. They also have a covariance function denoted as $G_l(t,t') = Cov(\psi_l(t), \psi_l(t'))$ which is a smooth function defined on $T \times T$. Using Mercer's Theorem, we can decompose it into the following:

$$G_l(t,t') = \sum_{m \geq 1} \kappa_{lm} \xi_{lm}(t) \xi_{lm}(t'), \tag{4.5}$$

where $\kappa_{k1} \geq \kappa_{k2} \geq \cdots \geq 0$ and $\{\xi_{lm}(t)\}$ form an orthonormal basis in L^2. By the Karhunen-Loève theorem, we get the expression:

$$\psi_l(t) = \sum_{m=1}^{\infty} \eta_{lm} \xi_{lm}(t), \tag{4.6}$$

where $\eta_{lm} = \int \psi_l(t) \xi_{lm}(t) dt$ are random variables uncorrelated over m with zero mean and variance equal to κ_{lm}. Putting this all together, we can represent $U(s,t)$ as:

$$U(s,t) = \sum_{l=1}^{\infty} \psi_l(t) \phi_l(s) = \sum_{l=1}^{\infty} \sum_{m=1}^{\infty} \eta_{lm} \xi_{lm}(t) \phi_l(s). \tag{4.7}$$

Using the above decomposition, we define the population model:

$$X(s,t) = \mu(s,t) + U(s,t); \quad U(s,t) = \sum_{l=1}^{\infty} \psi_l(t) \phi_l(s), \tag{4.8}$$

where $\psi_l(t) = \sum_{m=1}^{\infty} \eta_{lm} \xi_{lm}(t)$ are the score functions and $\phi_l(s)$ are the time-invariant PCs.

In reality, one cannot observe a continuous function and instead the observations are discrete observations on a grid. Let $Y_{(j_1,j_2,j_3),k}$ be a random variable at voxel index (j_1, j_2, j_3) and time k, where $j_1 \in \{1, \ldots, J_1\}$, $j_2 \in \{1, \ldots, J_2\}$, $j_3 \in \{1, \ldots, J_3\}$, for $J_1, J_2, J_3 \in \mathbb{Z}$ and $k \in \{1, \ldots, K\}$ for $K \in \mathbb{Z}$. For convenience we will shorten the voxel indices to j and $\{J_1 \times J_2 \times J_3\} = J$ and for all $j, s_j \in S$. Finally, for all $k \in K, t_k \in \mathcal{T}$.

Denote the total number of voxels in Y as $\eta = J_1 \cdot J_2 \cdot J_3$.

We have n copies of Y_{jk} denoted as Y_{ijk}. These observations are discrete realisations of a random smooth processes $X_i(s,t)$ which are independent and identically distributed (iid) copies of $X(s,t) \in L^2(S \cup \mathcal{T})$. This defined the model

$$Y_{ijk} = X_i(s_j, t_k) + \varepsilon_{ijk}, \tag{4.9}$$

where $s_j \in S, t_k \in \mathcal{T}$ are s, t evaluated on a discrete grid of points. The noise of the i^{th} subject, denoted as ε_{ijk}, is independent and identically distributed (iid) with mean zero and variance σ^2_{ijk} at voxel j and time k. We assume the variance function of $X_i(s_j, t_k)$ is smooth and hence ε_{ijk} has a smooth variance in the neighbourhood of j.

Combining results from equation (4.8) with model (4.9) yields:

$$Y_{ijk} = \mu(s_j, t_k) + \sum_{l=1}^{\infty} \psi_{il}(t_k)\phi_l(s_j) + \varepsilon_{ijk}. \tag{4.10}$$

We assume that there exists a number of PCs, denoted L, that contains a sufficient amount of variance explained so that any remaining information can be considered noise. Hence, we simplify the infinite sum to propose the full model:

$$Y_{ijk} = \mu(s_j, t_k) + \sum_{l=1}^{L} \psi_{il}(t_k)\phi_l(s_j) + \varepsilon_{ijk}. \tag{4.11}$$

Note that despite similar notation, ε_{ijk} in equations (4.10) and (4.11) are different. The variance explained (VE) by the model is given by:

$$\mathrm{VE}(L) = 1 - \left[\frac{\sum_{i=1}^{n} \sum_{k=1}^{K} \sum_j \left[Y_{ijk} - \mu(s_j, t_k) - \sum_{l=1}^{L} \psi_{il}(t_k)\phi_l(s_j) \right]^2}{\sum_{i=1}^{n} \sum_{k=1}^{K} \sum_j \left[Y_{ijk} - \overline{Y}_{jk} \right]^2} \right] \tag{4.12}$$

Intuitively, the model can be interpreted as the principal components representing variation in space whilst the score functions relate the PCs to random observations and show temporal variation as well. In further analysis, the PCs can be used to recover regions most relevant to data variation and the scores can be used in further analysis such as regression or clustering as they contain subject-specific information.

4.3 Parameter Estimation and Application

4.3.1 Estimation Methods

Traditionally, a covariance matrix would be calculated from which the PCs can be derived. To compute such a matrix, for every pixel in the three-dimensional image at one time-point, a covariance needs to be computed between it and the entire image again, creating an 8 dimensional matrix. This poses a computational burden, and hence we propose a new estimation method to avoid it outlined below.

We want to estimate $\psi_{il}(t)$ and $\phi_l(s)$, the principal components and score functions. To do this, we first estimate the mean across subjects denoted by $\mu(s,t)$, and the subject specific mean and variance denoted by $\mu_i(s,t)$ and $\sigma_i^2(s,t)$. The scores $\psi_{il}(t)$ are estimated by singular-value-decomposition and then the PCs $\phi_l(s)$ are estimated using regression. The remaining subsections will go over each step in detail, but briefly, the steps of the estimation algorithm are:

1. Per time point k, estimate the mean $\hat{\mu}_{jk}$ and $\hat{\sigma}^2_{ijk}$.
2. Estimate the raw score functions $\tilde{\psi}_{il}(t_k)$. This is done per time-point k using the eigendecomposition of the matrix made up of the inner product of Y_{ijk} with $\hat{\sigma}^2_{ijk}$ removed.
3. Regress the discrete data Y_{ijk} on $\tilde{\psi}_{il}(t_k)$ to obtain an estimate of the components $\tilde{\phi}_l(s)$, smooth it over s to obtain $\hat{\phi}_l(s)$.
4. Regress Y_{ijk} on $\hat{\phi}_l(s)$ to update the loadings, denote the updated loadings as $\hat{\psi}_{il}(t)$.

Note that steps 1 and 2 are done only once, whilst steps 3 and 4 can be repeated iteratively to improve the estimation of the components and scores. The implementation of this algorithm is available within the `Spatio_Temporal_FPCA` repository on GitHub.

Estimating the mean and variance

The estimation of the mean $\mu(s,t)$ is done by taking the average across subjects of all voxels at all time points, using the matrix denoted $\hat{\mu}_{jk} \in \mathbb{M}_{J,\mathcal{T}}$ where each element $\hat{\mu}_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_{ijk}$. To estimate the variance $\sigma_i^2(s,t)$ of ε_{ijk} from equations (4.9) and

(4.10), we exploit the smoothness property of $X_i(s,t)$ which assumes X to have little variance in the neighbourhood of s and hence any variance found in the neighbourhood of j in Y_{ijk} is predominantly attributed to the noise.

At each time point k, partition Y_{ik} for subject i into $h \times h \times h$ cubes, denoted with $m_{j'}$ where $j' = (j'_1, j'_2, j'_3)$ and $j'_1 \in \{1, \dots, \lfloor S_1/h \rfloor\}$, $j'_2 \in \{1, \dots, \lfloor S_2/h \rfloor\}$, and $j'_3 \in \{1, \dots, \lfloor S_3/h \rfloor\}$. For each $m_{j'}$ there are h^3 points j such that $Y_{ijk} \in m_{j'}$. Then we estimate variance $\sigma_i^2(s,t)$ in neighbourhood j' as $\hat{\sigma}^2_{ij'k} = Var(Y_{ijk})$ with $Y_{ijk} \in m_{j'}$.

Given estimate of $\hat{\sigma}^2_{ijk}$, construct diagonal matrices $W_{\sigma k}, k \in K$ with dimensions $n \times n$, where the $(i,i)^{th}$ entry is equivalent to $\int_S \sigma_i^2(s,t)ds$. As we have estimated σ^2, each diagonal entry is the discretization of the integral: $\sum_j \hat{\sigma}^2_{ijk}/\eta$, where η is the total number of voxels in one image.

Estimating Score functions

To estimate the score functions, we use the fact that the inner product of $X(\cdot)$ can be written as a function of the scores $\psi_l(t_k)$. Since Y_{ijk} is a function of $X(\cdot)$, specifically $Y_{ijk} = X_i(s_j, t_k) + \varepsilon_{ijk}$ we can use this alongside the representation of $X(\cdot)$ for the estimation procedure.

Define the matrix $W_{Xk} \in \mathbb{M}_{n,n}$ and $W_{Yk} \in \mathbb{M}_{n,n}$ with their $(i,i')^{th}$ entries denoted as $W_{Xk(i,i')}$ and $W_{Yk(i,i')}$ as follows:

$$W_{Yk(i,i')} = \langle Y_{ijk} - \hat{\mu}_{jk}, Y_{i'jk} - \hat{\mu}_{jk} \rangle = 1/\eta \sum_j \big(Y_{ijk} - \hat{\mu}_{jk}\big) \cdot \big(Y_{i'jk} - \hat{\mu}_{jk}\big),$$
$$W_{Xk(i,i')} = \langle X_i(\cdot,k) - \mu(\cdot,k), X_{i'}(\cdot,k) - \mu(\cdot,k) \rangle = \int_S U_i(s,k) U_{i'}(s,k) ds.$$

Note that due to the fact that $U_i(s,k)$, is a function of the scores, $W_{Xk(i,i')}$ can also be written as: $W_{Xk(i,i')} = \sum_{l=1}^{\infty} \psi_{il}(t_k) \cdot \psi_{i'l}(t_k)$, by using the properties of the PCs. Given this, the score functions $\psi_{ij}(t)$ can be estimated at the observed time points k. However, as only discrete data Y_{ijk} have been observed, W_{Xk} will be estimated using W_{Yk} and the population model (4.9).

$$\begin{aligned} W_{Yk(i,i')} &= \frac{1}{\eta}\sum_J \big(Y_{ijk} - \hat{\mu}_{jk}\big)\cdot\big(Y_{i'jk} - \hat{\mu}_{jk}\big) \\ &= \frac{1}{\eta}\sum_J \Big(\big(\sum_{l=1}^{\infty}\psi_{il}(t_k)\phi_l(s_j) + \varepsilon_{ijk}\big)\cdot\big(\sum_{l=1}^{\infty}\psi_{i'l}(t_k)\phi_l(s_j) + \varepsilon_{i'jk}\big)\Big) \\ &= \frac{1}{\eta}\sum_{l=1}^{\infty}\psi_{il}(t_k)\cdot\psi_{i'l}(t_k) + \frac{1}{\eta}\sum_J \varepsilon_{ijk}\cdot\varepsilon_{i'jk} \\ &\approx \sum_{l=1}^{\infty}\psi_{il}(t)\cdot\psi_{i'l}(t) + \frac{1}{\eta}\sum_J \varepsilon_{ijk}\cdot\varepsilon_{i'jk} \end{aligned}$$

Therefore, we can derive the following:

$$W_{Yk(i,i')} \approx \begin{cases} W_{Xk(i,i)} + \dfrac{1}{\eta}\displaystyle\sum_j \hat{\sigma}^2_{ijk} & \text{if } i = i' \\ W_{Xk(i,i')} & \text{otherwise,} \end{cases}$$

which can be shortened to $W_{Yk} \approx W_{Xk} + W_{\sigma k}$.

For matrix notation, define $\Psi_k \in \mathbb{M}_{n,L}$, where its il^{th} element is $\psi_{il}(t_k)$. Then the above notation can be rewritten as $W_{Yk} = \Psi_k\Psi_k^T + W_{\sigma k}$. The score matrix Ψ_k can be computed using the eigendecomposition of the matrix $W_{Xk} \approx W_{Yk} - W_{\sigma k} = \hat{R}\hat{P}\hat{R}^T$ where R is the matrix where each column is an eigenvector of $W_{Yk} - W_{\sigma k}$, and P is a diagonal matrix containing the corresponding eigenvalues. Then

$$\tilde{\Psi}_k = \hat{R}\hat{P}^{1/2}, \tag{4.13}$$

where each row of $\tilde{\Psi}_k$ corresponds to $\tilde{\psi}_{il}(t_k)$ and given all K matrices they form the set of first-estimate functions $\tilde{\psi}_{il}(\cdot)$ at observed time points.

Estimating and Smoothing PCs

The PCs correspond to the relationship between the observed Y_{ijk} and the estimated $\tilde{\psi}_{il}(t_k)$. This relation will be estimated via regression of the scores over the image, however as the images are high dimensional, we will vectorise Y_{ijk} per voxel j. The PCs will be the estimated regression coefficients.

Define $V_{ijk} = Y_{ijk} - \hat{\mu}_{jk}$ to be the observations with the global mean removed. We then vectorise V_{ijk} and $\tilde{\psi}_{il}(t_k)$. For each $j \in J$, let

$$vec(V_j) = \left[V_{1j1}, V_{2j1}, ..., V_{1j2}, ..., V_{njK}\right],$$

be the vector form of the data observed for all subjects and all time points at point j. Similarly, form a vector of the estimated score functions

$$vec(\tilde{\psi}_l) = \left[\tilde{\psi}_{1l}(t_1), \tilde{\psi}_{2l}(t_1), ..., \tilde{\psi}_{1l}(t_2), ..., \tilde{\psi}_{nl}(t_K)\right].$$

The design matrix $\tilde{\Psi} = [1, vec(\tilde{\psi}_1), ..., vec(\tilde{\psi}_L)]$ is formed from each of the vectorised score function where 1 denotes a vector of ones. This can be used to form the simple linear regression:

$$vec(V_j) = \beta_{j0} + \beta_{j1} vec(\tilde{\psi}_1) + \cdots + \beta_{jL} vec(\tilde{\psi}_L) + \epsilon_j.$$

Estimating the coefficients is done using least squares, $\hat{\beta}_j = \left(\tilde{\Psi}^T\tilde{\Psi}\right)^{-1}\tilde{\Psi}^T vec(Y_j)$, where $\hat{\beta}_j = \left[\beta_{j0}, \beta_{j1}, \ldots, \beta_{jL}\right]$. The above process is a point-by-point estimation of a functional regression model which would be written as follows:

$$V_{ijk} = \beta_0(s_j) + \sum_{l=1}^{L} \beta_l(s_j)\tilde{\psi}_{il}(t_k) + \varepsilon_{ijk} \tag{4.14}$$

where $\beta_l(s_j)$ is the collection of the point-wise regression coefficient β_{jl}. It follows that for each j, $\tilde{\phi}_l(s_j) = \beta_{jl}$.

The first estimated PCs $\tilde{\phi}_l(s_j)$ are subsequently smoothed using a sandwich smoothing method first introduced by Xiao et al., 2012. This approach applies smoothing matrices S_1 and S_2 to some matrix H to obtain a smooth matrix $\hat{H}$: $\hat{H} = S_1 H S_2$. Each smoothing matrix S_i is a univariate matrix constructed using b-splines and differencing matrices which will be defined later. The above equation can be rewritten using tensor products and their properties so that it becomes

$$\hat{H} = (S_1 \otimes S_2)H.$$

This equation makes it simple to extend the smoothing method to any p-dimensional matrix, as described in Section 7 of Xiao et al., 2012.

To smooth $\tilde{\phi}_l(s_j)$ we define 3 smoothing matrices where each matrix S_i is constructed as follows for $i = 1, 2, 3$:

$$S_i = B_i(B_i^T B_i + \gamma_i D_i^T D_i)^{-1} B_i^T.$$

Here, each B_i are the model matrices defined on each of the x, y, z dimensions using B-spline basis vectors defined over each domain. D_i is the differencing matrix, a matrix representation of the difference operator Δ, which acts on some sequence a_j, such that $\Delta a_j = a_j - a_{j-1}$. Both the construction of D_i and B_i, and their use in smoothing, are nicely explained in Eilers and Marx, 1996. The coefficient γ is added for continuous control over smoothness of fit.

Given the three smoothing matrices S_1, S_2 and S_3 and a matrix representation of $\tilde{\phi}_l(s_j)$ denoted as $\tilde{\phi}_l$, the matrix estimate of the principal component is obtained as follows

$$\hat{\phi}_l = (S_3 \otimes S_2 \otimes S_1)\tilde{\phi}_l(s_j).$$

Functional Regression of Images on the PCs

The estimated PCs will be used to create an updated version of the score functions. So far, we have treated the estimated score function as discrete observations on a continuum. We would like to update the score functions given the new estimated PCs and smooth them to have a continuous function over time. This is done by functional regression of V_{ijk} on $\hat{\phi}_l(s)$ to update the scores, which uses basis functions to create estimated coefficient functions. We denote the updated scores as $\hat{\psi}_{il}(t)$.

Given V_{ijk} as before, we want to find functions $\beta_{il}(t)$ such that

$$V_{ijk} = \sum_{l=1}^{L} \beta_{il}(t_k)\hat{\phi}_l(s_j) + \varepsilon_{ijk},$$

where ε_{ijk} is noise following $\mathcal{N}(0, \tilde{\sigma}^2_{ijk})$. We use tilde to separate $\tilde{\sigma}^2$ from our estimated variance $\hat{\sigma}^2$ from Section 4.3.1. Note that this process is done separately for each subject as denoted with the index i.

As we cannot directly estimate $\beta_{il}(t)$, this functional regression problem will be reworded. Suppose $\{\zeta_1(\cdot), ..., \zeta_{\mathcal{K}}(\cdot)\}$ is a set of pre-specified basis functions. The coefficient functions can be expanded as:

$$\beta_{il}(t) = \sum_{\kappa=1}^{\mathcal{K}} b_{il\kappa}\zeta_{i\kappa}(\cdot).$$

Then, the regression model from before can be expressed as:

$$V_{ijk} = \sum_{l=1}^{L} \phi_l(s)\big(\sum_{\kappa=1}^{\mathcal{K}} b_{il\kappa}\zeta_{i\kappa}(t)\big) + \varepsilon_{ijk}, \tag{4.15}$$

which thus reduces the problem to estimating the coefficients $\{b_{ilk}\}$ where $l \in \{0, ..., L\}, \kappa \in \{1, ..., \mathcal{K}\}$.

In practice, functions are observed on a discrete grid and hence we can represent the above as a set of matrices. Let V_i be an $S_1 \times S_2 \times S_3 \times \mathcal{T}$ array of the i^{th} subject's voxels with the mean $\hat{\mu}_{jk}$ removed. Define Φ_l as an array of the function $\phi_j(s_j)$ evaluated on the same grid of points j_1, j_2, j_3 where where $j_1 \in [0, S_1] \cap \mathbb{Z}$, $j_2 \in [0, S_2] \cap \mathbb{Z}$, $j_3 \in [0, S_3] \cap \mathbb{Z}$. Let Z be the $K \times \mathcal{K}$ matrix whose columns correspond to the K basis functions $\zeta_{ik}(t_\tau)$, for $\tau \in \mathcal{T}$, where t_τ denotes the function is evaluated at discrete time points. Finally, let B_i be the $L \times \mathcal{K}$ matrix with jth row being the vector of basis coefficients for $b_{il\kappa}$ (the first coefficient is omitted as V_{ijk} is centered so we don't need $(L+1)$). Then equation (4.15) can be expressed as:

$$V_i = \sum_l \Phi_l B_i Z^T + E_i, \tag{4.16}$$

where E_i is the $S_1 \times S_2 \times S_3 \times \mathcal{T}$ array of error terms. This model can be posed as a standard linear model. Let $vec(V_i^T)$ be the vector formed by concatenating the rows of U_i, and note that

$$vec\big((\Phi_l B_i Z^T)^T\big) = \big(\Phi_l \otimes Z\big)vec\big(B_i^T\big),$$

where $\otimes$ represents the kronecker product of two matrices. Then the regression problem takes the form:

$$vec\big(V_i\big) = \big(\Phi_l \otimes Z\big)vec\big(B_i^T\big) + vec\big(E_i^T\big), \tag{4.17}$$

and the coefficients B can be estimated using least squares from $vec\big(B_i^T\big)$.

Finally, steps described in Sections (4.3.1) and (4.3.1) can be repeated to improve the estimated PCs and score functions.

4.4 Simulation

The goal of the simulation is to evaluate our proposed estimation method in identifying the true number of PCs, L, ability to recover score and PC functions and reconstructing the data. To understand the effect of noise on the estimation performance, we considered various noise settings. For fitting we use the true model and consider two data scenarios: one where the simulated data agrees with the fitted model and one where it does not.

4.4.1 Design

The simulation designs vary in complexity. We first simulate two different scenarios where the data generated adheres to the fitted model (Designs 1 and 2). In addition, we fit $L = 3$ components when the underlying model is set to $L = 2$. We also run two simulations where the data is generated with a model that deviates from the one that is fitted. In this setting one design matches complexity from the models described in this chapter whilst the second simulation increases the complexity. In all designs, we assume the global mean $\mu(s, t) = 0$ and we set the principal component number $L = 2$.

Simple Simulation (Design 1)

Define:

$$\psi_{i1}(t) = a_i \cdot \cos(0.5\pi t), \qquad \psi_{i2}(t) = b_i \cdot \sin(\pi t)$$
$$\phi_1(s) = \phi_1(s_1, s_2, s_3) = \sqrt{2} \cdot \cos(2\pi s_1), \qquad \phi_2(s) = \phi_2(s_1, s_2, s_3) = \sqrt{2} \cdot \sin(2\pi s_1),$$

where $t \in [0, 1]$, $s \in [0, 1] \times [0, 1] \times [0, 1]$, $a_i \sim N(0, 2)$ and $b_i \sim N(0, 0.5)$. The score functions are evaluated on a grid of 20 equidistant time points denoted $t_k \in [0, 1]$, specifically $t_k \in \{\frac{1}{20}, \ldots, 1\}$. The functions ϕ_l are evaluated on a $30 \times 30 \times 30$ grid, where $s_{j1}, s_{j2}, s_{j3} \in \{\frac{1}{30}, \ldots, 1\}$, a point on this grid will be denoted s_j. The functions form simulated images on a $30 \times 30 \times 30$ grid over 20 time points using the model equation

(4.11):

$$Y_i(s_j, t_k) = \sum_{l=1}^{2} \psi_{il}(t_k)\phi_l(s_j) + \varepsilon_{ijk}, \tag{4.18}$$

where $i \in \{1, \ldots, n\}$ for $n = 100$ and ε_{ijk} follows one of the distributions from the three noise settings defined below.

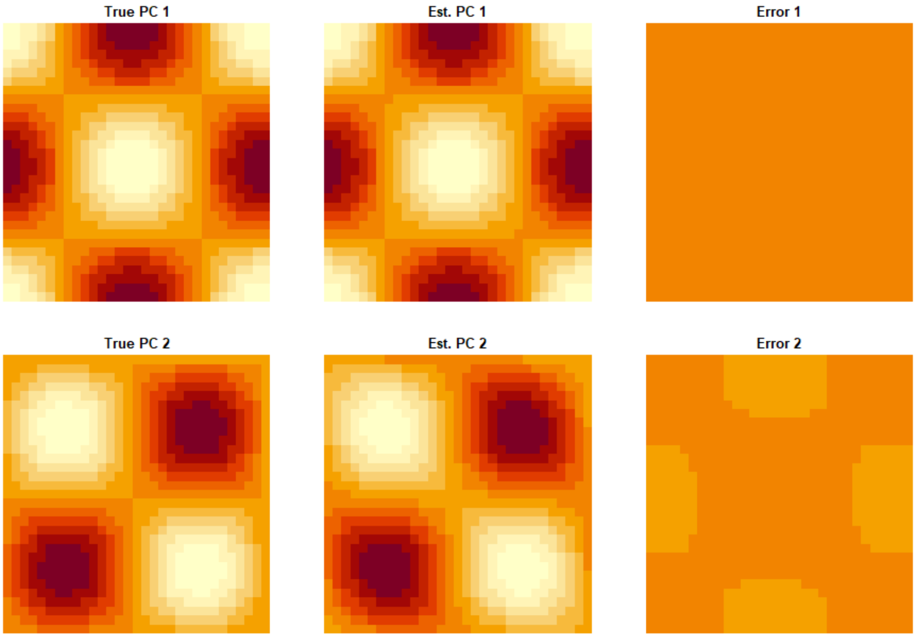


Figure 4.1: True and estimated PCs from one replicate of a simulation with design 2 (no noise and $L = 2$). The image represents a slice along the z-axis at $z = 15$ of a 3-dimensional object. The error is the difference between the True and Estimated PCs along the slice.

Complex Simulation (Design 2)

Define:

$$\psi_{i1}(t) = a_i \cdot \cos(b_i \cdot 0.5\pi t), \qquad \psi_{i2}(t) = c_i \cdot \sin(d_i \cdot \pi t)$$

$$\phi_1(s) = \phi_1(s_1, s_2, s_3) = \sqrt{2}^3 \cdot \cos(\pi s_1) \cdot \cos(\pi s_2) \cdot \cos(\pi s_3),$$

$$\phi_2(s) = \phi_2(s_1, s_2, s_3) = \sqrt{2}^3 \cdot \sin(\pi s_1) \cdot \sin(\pi s_2) \cdot \sin(\pi s_3),$$

$$a_i \sim N(0, 2), \quad b_i \sim N(0.85, 0.25), \quad c_i \sim N(0, 0.5), \quad d_i \sim N(1, 0.5)$$

where $t \in [0, 1]$, $s \in [0, 1] \times [0, 1] \times [0, 1]$. The grid of points for t_k and s_j is defined as in Design 1. Images are generated as in eqn. (4.18) and the sample size is set to $n = 100$.

The Design 1 and 2 simulation performance will be evaluated with each of the following noise settings:

1. $\varepsilon_{ijk} = 0$,
2. $\varepsilon_{ijk} \sim N(0, 0.1Var(Y_{ijk}))$,
3. $\varepsilon_{ijk} \sim N(0, 0.2Var(Y_{ijk}))$.

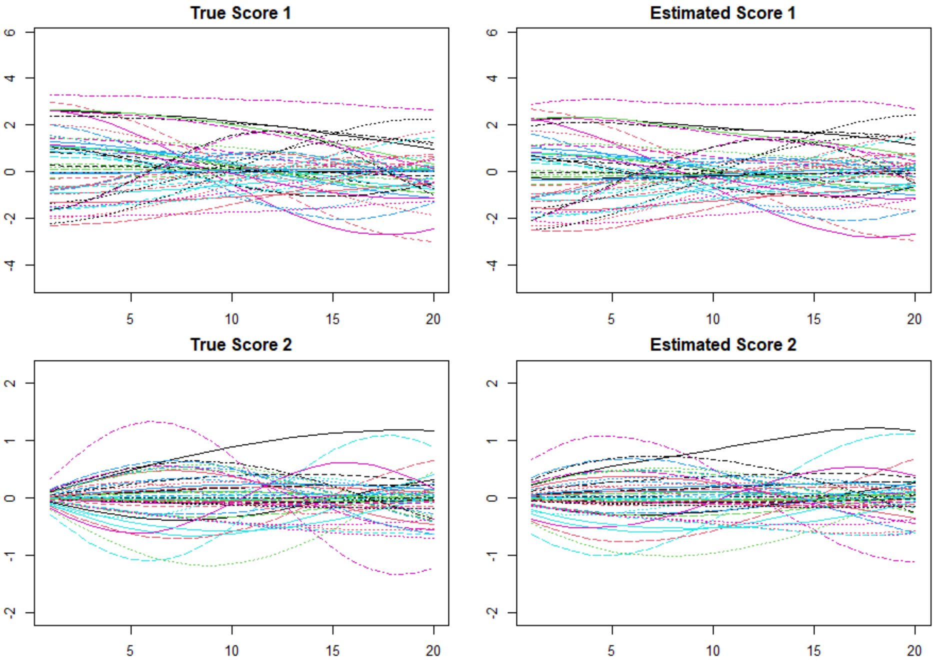


Figure 4.2: True and estimated score functions from one replicate (Design 2). Each line corresponds to the i^{th} simulated image generated.

Interactions Between Space And Time (Designs 3 and 4)

In the following designs, we let the PCs depend on time as well as space. In contrast to previous designs where scores and PCs were on completely separate domains, we would like to understand the effect of more space-time interactions on the model.

For Designs 3 and 4, the scores and principal components will change their domain to include or exclude time, which will cause the data to deviate from the underlying model assumptions. In both designs, the principal components are defined to include time dependence as follows:

$$\phi_1(s) = \phi_1(s_1, s_2, s_3) = \sqrt{2}^3 \cdot \cos(\pi s_1) \cdot \cos(\pi s_2) \cdot \cos(\pi s_3),$$
$$\phi_2(s,t) = \phi_2(s_1, s_2, s_3, t) = \begin{cases} \sqrt{2}^3 \cdot \sin(\pi s_1) \cdot \sin(\pi s_2) \cdot \sin(\pi s_3) & \text{if } s_1 > 0.5, t > 0.5 \\ 0 \quad \text{otherwise,} \end{cases}$$

where $s \in [0,1] \times [0,1] \times [0,1]$ and $t \in [0,1]$. These are displayed in Fig. (4.3). For Design 3, the first score will follow design 2 whilst the second one will be a scalar:

$$\psi_{i1}(t) = a_i \cdot \cos(b_i \cdot 0.5\pi t), \qquad \psi_{i2} = c_i$$
$$a_i \sim N(0,2), \quad b_i \sim N(0.85, 0.25), \quad c_i \sim N(0, 0.65),$$

where t is defined as above. For Design 4 both scores will be functions over time:

$$\psi_{i1}(t) = a_i \cdot \cos(b_i \cdot 0.5\pi t), \qquad \psi_{i2}(t) = c_i \cdot \sin(d_i \cdot \pi t)$$
$$a_i \sim N(0,2), \quad b_i \sim N(0.85, 0.25), \quad c_i \sim N(0, 0.5), \quad d_i \sim N(1, 0.5)$$

where $t \in [0,1]$. The grid of points t_k and s_j is defined as before. For $\phi_2(s,t)$, this means it is equal to zero for $s_{j_3} < 15$ and $t_k < 10$.

Data for Design 3 and 4 are generated using:

$$D3 : Y_i(s_j, t_k) = \psi_{i1}(t_k)\phi_1(s_j) + \psi_{i2} \cdot \phi_2(s_j, t_k) + \varepsilon_{ijk},$$
$$D4 : Y_i(s_j, t_k) = \psi_{i1}(t_k)\phi_1(s_j) + \psi_{i2}(t_k) \cdot \phi_2(s_j, t_k) + \varepsilon_{ijk}.$$

In both cases we set $n = 100$ and the noise is set to $\varepsilon_{ijk} \sim N(0,\ 0.1Var(Y_{ijk}))$.

Models Fitted

For all designs, we fit the proposed model with $L = 2$. For D1 and D2 specifically, we consider fitting an additional component. The score functions are estimated using function-on-function regression described in Section 4.3.1. To avoid over-fitting whilst still preserving the shape of the underlying functions, the scores are represented using 4 b-spline vectors of order 3.

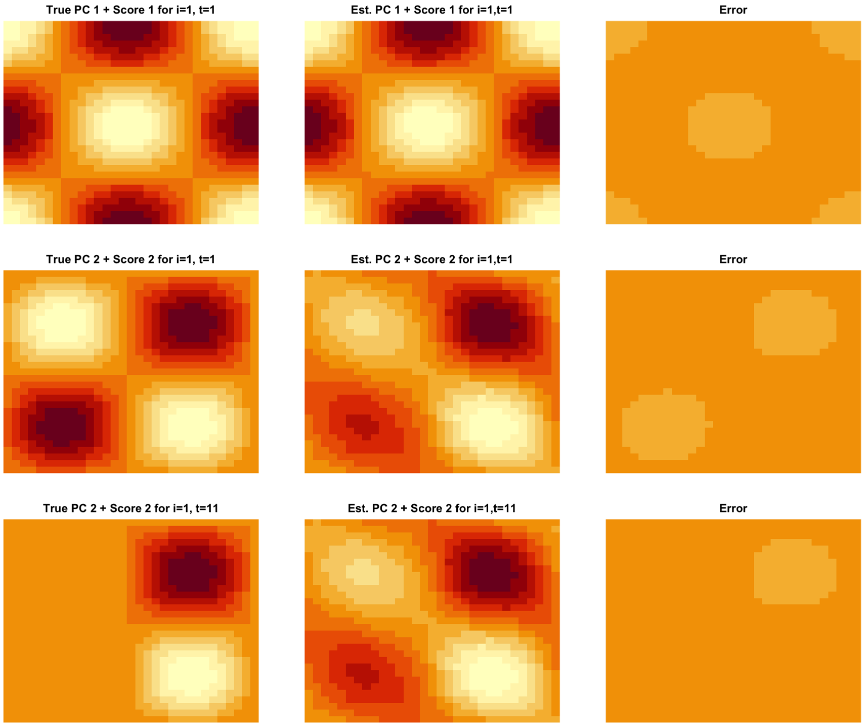


Figure 4.3: Principal Components for Designs 3 and 4 along $s_{j_3} = 15$ slice.

Evaluation

Estimation accuracy for the model components is evaluated using integrated square error (ISE) for the PCS and mean integrated squared error (MISE) for the scores. They are defined as follows:

$$\text{ISE}(\hat{\phi}_l(s)) = \int (\phi_l(s) - \hat{\phi}_l(s))^2 \, ds$$

$$\text{MISE}(\hat{\psi}_l(t)) = \frac{1}{n} \sum_{i=1}^{n} \big[\int (\psi_{il}(t) - \hat{\psi}_{il}(t))^2 \, dt \big].$$

This will be done per replicate. Over the course of the full simulation, the mean and standard deviation of the above errors will be used to summarize the estimation performance across all the replicates. The overall reconstruction is evaluated using variance explained (VE) from equation (4.12). In simulations where noise is present, we will also compute VE where the estimated image $\hat{Y}_i(s_j, t_k)$ (reconstructed from parameters that were estimated with noisy image as input) is compared to $Y_i(s_j, t_k)$ without the addition of noise. This will be referred as 'VE Clean' (VEC).

4.4.2 Simulation Results

Summary of Results for Design 1 and 2

The reconstruction error and VE for each simulation design are summarised in Table (4.1). Table (4.2) shows the mean and the standard deviation of the ISE and MISE of the estimated score functions and PCs across the replicates.

In Design 1, when no noise is present, the mean VE from a two PC reconstruction is 0.96(0.007). The total estimated VE decreases with increasing noise. VEC is 0.96 for both noise settings of 10% and 20%. Looking at VE by individual PCs, estimated PC1 explains less VE (0.642) than by design, whilst PC2 explains more (0.318). This VE ratio between the PCs did not change significantly with noise. Concerning the errors (Table 4.2), the PCs $\hat{\phi}_l(s)$ and score functions $\hat{\psi}_{il}(t)$ had a consistent error over the course of the simulations, and the MISE was not affected by different noise settings.

Noise	**RE**	**Variance Explained**										
		True			Estimated							
		$\phi_1(\cdot)$	$\phi_2(\cdot)$	Total	$\hat{\phi}_1(\cdot)$		$\hat{\phi}_2(\cdot)$		Total		Clean	
					mean	sd.	mean	sd.	mean	sd.	mean	sd.
D 1												
0 %	0.878	0.790	0.210	1.000	0.642	0.033	0.318	0.029	0.960	0.007	na	na
10 %	1.079	0.719	0.191	0.910	0.619	0.034	0.303	0.022	0.922	0.009	0.960	0.005
20 %	1.687	0.658	0.175	0.833	0.580	0.033	0.280	0.028	0.860	0.014	0.958	0.006
D 2												
0 %	0.862	0.800	0.200	1.000	0.656	0.060	0.305	0.034	0.961	0.054	na	na
10 %	2.387	0.727	0.182	0.910	0.619	0.030	0.281	0.028	0.905	0.086	0.927	0.007
20 %	4.362	0.667	0.167	0.834	0.589	0.033	0.270	0.029	0.859	0.125	0.925	0.010

Table 4.1: Average reconstruction reconstruction error (RE) and variance explained over 100 replicates. True VE is the proportion of variance the PC explains given a level of noise in the data. VEC is comparing the reconstructed image to a true image with no noise. D1 and D2 refer to Design 1 and 2 respectively.

In Design 2 the average total VE is 0.96 with no noise in the data. Total VE decreases with increasing noise, and the estimated sd. increases from 0.5 to 0.13. VEC is 0.93 for both noise levels with sd estimated to 0.01. Similar to D1, estimated principal components don't contribute the same level of VE as per design. PC1 explains 0.656(0.060) and PC2 explains 0.305(0.034), these values decrease with noise but the ratio remains similar as in D1. The average MISE over the replicates shown in Table (4.2) shows small errors for simulation with no noise, but in all function error metrics, the error increases significantly with the presence of noise in the data.

	Error Measurements over All Replicates							
	$\psi_1(t)$		$\psi_2(t)$		$\phi_1(s)$		$\phi_2(s)$	
D 1	mean	sd.	mean	sd.	mean	sd.	mean	sd.
0 %	0.166	0.042	0.149	0.050	0.087	0.051	0.174	0.093
10 %	0.161	0.040	0.148	0.048	0.078	0.047	0.158	0.062
20 %	0.164	0.035	0.155	0.056	0.082	0.057	0.165	0.108
D 2								
0 %	0.138	0.055	0.109	0.030	0.036	0.050	0.047	0.052
10 %	0.339	0.028	0.236	0.025	0.313	0.012	0.320	0.019
20 %	0.340	0.036	0.238	0.025	0.314	0.013	0.322	0.021

Table 4.2: The mean and standard deviation of MISE and ISE for the score functions and PCs, respectively, over 100 replicates.

Figures (4.1) and (4.2) show the D2 estimated functions next to the true ones. Whilst the estimated shape for the PCs appears to be well estimated, the score functions appear to have different local minima and maxima over the domain. Furthermore, $\hat{\psi}_2(t)$ at $t = 0$ could not match the range of the true. Overall, the shapes were preserved in the estimation, however exact function qualities have been changed.

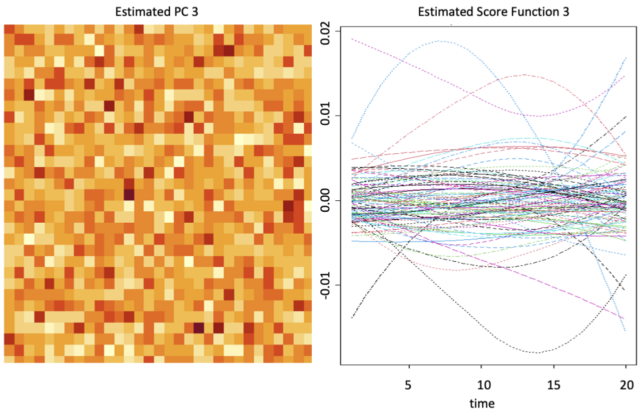


Figure 4.4: An example of an estimated 3rd PC and score function from one replicate (design 2 with noise 20%).

Maximum of Principal Components

We first wanted to understand the estimation methods per se, next we would like to evaluate the proposed methods ability to chose an appropriate number of components to estimate. To do this, in simulation designs 1 and 2, we attempted to estimate 3 PCs when then underlying number is 2. When no noise is present, the maximum level of possible estimated PCs is limited to 2. This is due to first estimate of score functions estimating a negative eigenvalue in the matrix R from eq. (4.13) when attempting to evaluate the $(L+1)^{th}$ eigenfunction. Given this, only L score functions could be evaluated. In the presence of noise the $(L+1)^{th}$ eigenfunction in eq. (4.13) could be evaluated and hence 3rd score and PC functions were obtained. An example of such functions can be seen in Fig. (4.4). Indeed, the third set of functions resembles noise. The scores are centered around zero in the range [-0.01, 0.01] whilst the PC is a three-dimensional matrix of noise. When added to VE these components provide no visible improvement or reduction in estimated reconstruction of the data.

Simulation Results for Design 3 and 4

Table (4.3) shows the reconstruction error from the estimated model components in Designs where the data is generated using components that don't satisfy our model assumptions. For D3, the total VE is relatively high at 0.93(0.05) and VEC 0.95(0.05). For D4, with a more complex design, the VE is 0.74(0.02) and VEC is 0.90(0.03).

Noise	**RE**	**Variance Explained**										
		True			Estimated							
		$\phi_1(\cdot)$	$\phi_2(\cdot)$	Total	$\hat{\phi}_1(\cdot)$		$\hat{\phi}_2(\cdot)$		Total		Clean	
					mean	sd.	mean	sd.	mean	sd.	mean	sd.
D3: 10%	2.86	0.65	0.25	0.91	0.59	0.04	0.35	0.02	0.93	0.05	0.95	0.05
D4: 10%	1.34	0.79	0.21	0.91	0.63	0.04	0.11	0.02	0.74	0.02	0.90	0.03

Table 4.3: Average reconstruction error (RE) and variance explained over 100 replicates. True VE is the proportion of variance the PC explains given a level of noise in the data. VEC is comparing the reconstructed image to a true image with no noise. D3 and D4 refer to Design 3 and 4 respectively.

Table 4.4 depicts the MISE of the score and PC functions. Beginning with the scores, to compare ϕ_2 to $\hat{\phi}_2$ we treat ϕ_2 as a constant function over time. Overall, the error in both designs is comparable with $\psi_1(t)$ in D4 having the largest average error over all replicates. In all cases the sd for these metrics is low. Comparing these results with D2 (as the score function is the same), we find the average error and sd to be similar.

Looking at the PCs, the error for $\phi_1(s)$ in both D3 and D4 is comparable to that in D2. To estimate the error for $\phi_2(s,\cdot)$ we consider three values, two errors are estimated at

	Error Measurements over All Replicates											
	$\psi_1(t)$		$\psi_2(t)$		$\phi_1(s)$		$\phi_2(s,1)$		$\phi_2(s,11)$		$\phi_2(s,t)$	
	mean	sd.	mean	sd.	mean	sd.	mean	sd.	mean	sd.	mean	sd.
D3: 10%	0.32	0.04	0.26	0.05	0.34	0.09	0.49	0.14	0.42	0.16	0.46	0.15
D4: 10%	0.39	0.03	0.24	0.04	0.32	0.047	0.46	0.07	0.38	0.05	0.43	0.06

Table 4.4: The mean and standard deviation of error metrics for the score functions and PCs over 100 replicates. ISE for $\phi_1(s)$, $\phi_2(s,1)$ and $\phi_2(s,11)$ integrate over s and ISE for $\phi_2(s,t)$ integrate over s and t.

two different stages of the function, the the third integrates over time completely. The estimation error for PC2 is higher than for PC1 which is consistent with the fact that the model had to interpolate between the two states of PC2 over time. This can be seen in Figure (4.3). Overall the error for PC2 is high in both designs and much higher when compared to the error in D2.

Simulation Conclusion

Looking at simulations where the fitted model matches the underlying one, we find that overall, the estimation method can reconstruct the true functions well and consequently, reconstruct the original image accurately. Estimated total VE matched the true VE when accounting for noise and VEC remained above 0.90 for all but one simulation (D2 with high noise). In short, VEC for D1 was more consistent regardless of noise, unlike D2 where the noise lowered CVE to 0.92. We can infer that the complexity of the underlying true functions makes the model more susceptible to noise.

We found that the method had the ability to estimate the correct number of components in setting without noise. When noise was present, the extra components were estimated of noise itself. For the above reasons, it was easy to determine the underlying number of components regardless of the noise setting.

In simulations when the underlying model deviated from the fitted one, we find that the overall accuracy of the reconstruction decreases, and functions that specifically deviate from model assumptions have worse estimations. The simpler deviation (D3) had higher VE values than D4. This could be due to the fact that the temporal dependency in D3 is quite simple (the second component and score could be rewritten using an indicator function) whilst this dependency is less straightforward in D4 with both PC2 and the score function having temporal variation.

In all designs, we find that the overall shape of the score functions was preserved as is shown on Fig. (4.2). The choice of basis vectors in the functional regression has a high effect of estimation accuracy. The choice of b-spline vectors has a limitation which can be seen in Figure (4.2) where $\hat{\psi}_2(t)$ there is more variation at time 0. Other choices such as Fourier functions were considered, but they were not able to reproduce smooth curves over t and they would not reproduce the starting point for $\hat{\psi}_2(t)$ accurately. Finally, basis vectors limit the ability to fully replicate the shape of the function, and hence we see slight deviations in minima/maxima of the estimated functions compared to the true ones. Overall, VE was good in all designs except for D4, we can clearly identify the right number of components and the model functions were estimated well.

4.5 Data Analysis

4.5.1 Background

The data consists of brain fMRI images of 15 subjects, taken whilst they performed 81 financial decision-making tasks. Subjects were presented with a investment return stream and then asked to perform one of 4 task categories that were related to risk perception scoring and expected return prediction. The images were acquired every 2.5 s during the investigation, providing each subject with a time series of 1360 three-dimensional images representing their brain activity, an example of one subjects image can be seen in Figure (4.5). The images represent the Blood Oxygenation Level Dependent (BOLD) signals. Each patient also received a clinically-derived risk score at the end.

Our aim is to recover active brain areas found in previous literature to be associated with financial decision making and to use the score functions in a logistic regression model to quantify associations between these ROIs and subject risk prevalence. In addition, we will compare our approach to the ones described in Li et al., 2019 where they approached this task by fitting a standard FPCA model to one image without directly modelling the time component. Each subjects' fMRI series was divided by task and for each task, the first 3 images were taken and concatenated into one. That is, if $Y_{iqk} \in \mathbb{M}_{J,K}$ is the i^{th} subject's image for task q and time t_k, their model is estimated on the array: $\tilde{Y}_{iq} = (Y_{iq2} - Y_{iq3})/2 - Y_{iq1}$. This was done under the assumption that the most important activity occurs at the begging of the task and taking the difference of the first three images should account for this.

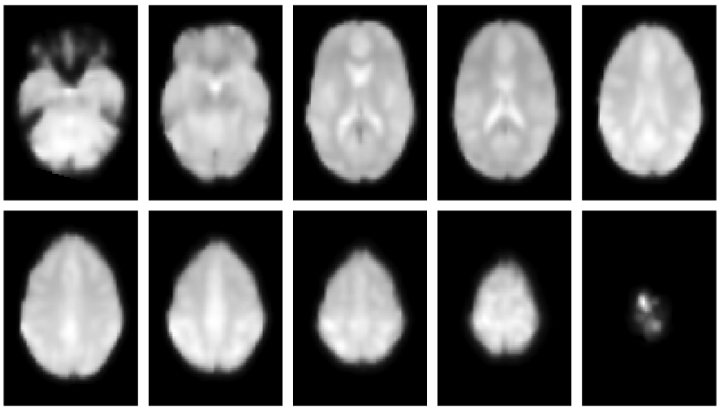


Figure 4.5: Raw data slices for one patient. This represents the data $Y_{1,(s_1,s_2,s_3),t}$ where $t = 1$, $s_1, s_2 \in [0, 91]$ and $s_3 \in \{25, 30, 40, 45, 50, 55, 60, 65, 70, 80)\}$.

Considering ROIs, previous studies found correlations between risk attitude and risk-related brain activity in the lateral orbitofrontal cortex (lOFC) for risk-averse individuals and in the medial orbitofrontal cortex (mOFC) for risk-seeking individuals (Tobler et al., 2007). Another study (Mohr et al., 2010b) found that inter-individual differences in decision-related brain activity in the lOFC and correlated it with inter-individual differences in risk attitudes independent of the current level of risk. The authors showed that the value signal in the ventrolateral prefrontal cortex (VLPFC) increased with risk in risk-seeking individuals and decreased with risk in risk-averse individuals, thereby reflecting the risk attitude. Additionally, Chen et al., 2015 have found the dorsolateral prefrontal cortex (DLPFC) and the anterior insula (aINS) to be correlated with decision making and risk.

4.5.2 Data Description and Pre-processing

The data is downloaded in a `.mat` format from the Humboldt University of Berlin website in December 2021. It is transformed into a 4D array of size 91 × 109 × 91 × 1360. Details of preprocessing can be found in section The data was initially pre-processed with FSL 4.0, which included motion correction, slice-time correction and spatial smoothing using a 8mm Gaussian kernel. Additionally, images were normalized into a standard space

(Mohr et al., 2010b). Time is divided by using the point at which each task has started (provided in milliseconds). Details of these procedures and the data itself are in section (3.3). We work under the assumption that the first image was taken at time equal to zero. Additionally, we only consider 15 images per task corresponding to 37.5 seconds, slightly longer than the 30 seconds it takes to complete (Mohr et al., 2010b and van Bömmel et al., 2013).

Data is first loaded into MATLAB as a complete series of three dimensional images for each subject. This is done to subdivide the series into sub-series connected to each task. Given that each task takes 30 seconds to complete, this corresponds to under 15 images. Assuming the image series starts at time 0 in seconds, we used the list of all task starting times (in milliseconds) to find the image corresponding to the start time, from there we chose a series of 15 images including the starting image. This creates 81 sub-series for each subject that correspond to each task.

The resulting sub-series are loaded into R. There are a few regions of the brain that have muted or blank voxel values which can be attributed to them being potentially outside the field of view, which is typically defined prior to the fMRI scan and should include all parts of the brain relevant for the task. As fMRIs tend to have a quick acquisition time, some areas outside of the field of view can be missed in the scans. We find missing voxels towards the bottom along the z-axis of the images. As the missingness is different across subject, the bottom 25 slices are removed. The resulting image is loaded into R as an array of dimensions $n = 60, S_1 = 83, S_2 = 100, S_3 = 67, \mathcal{T} = 15$.

Preliminary analysis showed that all tasks had the same order of PCs. However, during each task type the subjects would exhibit similar behaviour and hence including multiple tasks of the same type in the model would not provide more information. For this reason, we chose 4 different task types to be included in the data to be analysed. The task indices were chosen to be 1, 3, 23 and 26. The image arrays corresponding to these tasks were concatenated to create the full data array with $n = 15 \times 4$ subjects that can be denoted as Y_{iqjk} where $i = 15$ is the number of study participants, q is the index for the task, j is the index over three spatial dimensions and k corresponds to time.

	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_3(s)$	$\hat{\phi}_4(s)$	$\hat{\phi}_5(s)$	$\hat{\phi}_6(s)$	$\hat{\phi}_7(s)$
VE	0.130	0.116	0.088	0.083	0.072	0.072	0.071
	$\hat{\phi}_8(s)$	$\hat{\phi}_9(s)$	$\hat{\phi}_{10}(s)$	$\hat{\phi}_{11}(s)$	$\hat{\phi}_{12}(s)$	$\hat{\phi}_{13}(s)$	$\hat{\phi}_{14}(s)$
VE	0.051	0.044	0.041	0.037	0.037	0.031	0.028

Table 4.5: The Variance Explained of every estimated PC.

4.5.3 Implementation

We fit the model using the algorithm described in Section 4.3.1 to estimate 14 PCs and their corresponding score functions. Fourteen is the maximum number of components that could be estimated given the dataset, any following component estimated would appear to resemble noise. First the mean $\hat{\mu}_{j,k}$ is estimated by taking the average pixels across subjects. To estimate $\hat{\sigma}^2_{iqjk}$ cubes of size $3 \times 3 \times 3$ are chosen. When the PCs are smoothed using 3D penalized smoothing, we adopt the cubic b-spline basis with 30, 35, 30 knots along the x, y, z axes, respectively. This approach is similar to that used in Li et al., 2019. The penalty matrices for all directions are of order 2. Tuning parameters in the penalized smoothing are selected by minimizing the GCV values computed as done in Li et al., 2019. The scores are updated using functional regression with 10 b-spline basis vectors. The number of splines used to reconstruct the score functions was chosen to be as small as possible whilst still allowing for the representation of small variations without over-smoothing.

We are interested in two sets of PCs and score functions: one set is to see how well the model fits to the data and another set to identify the ROIs and their activity over time in subjects. One set, which will be denoted as $\hat{\phi}(s_j)$ contains the complete estimated PCs with the corresponding score functions and is evaluated in terms of reconstructing the image. In the other set, the PCs are trimmed by the 0.1% and 99.9% quantile levels for each component l such that for any voxel j in the quantile $\hat{\varphi}(s_j) = 1$, and $\hat{\varphi}(s_j) = 0$ otherwise.

Image Reconstruction

The complete estimated PCs and the score functions are used to reconstruct the original data. The overall reconstruction is evaluated using VE estimated using equation (4.12). Overall, using 14 PCs leads to VE of 90.1%.

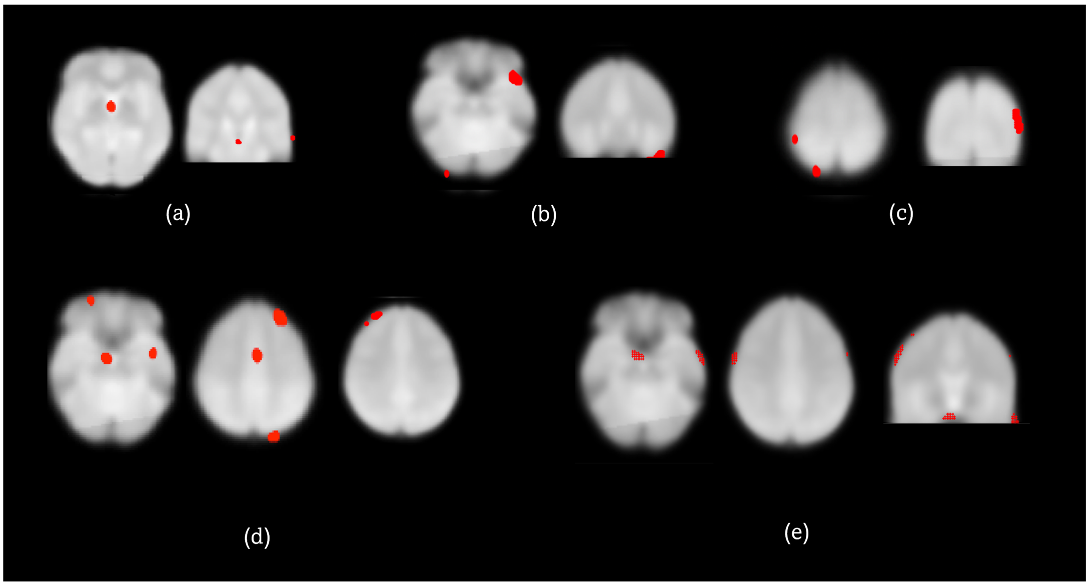


Figure 4.6: ROIs recovered from 5 different principal components: (a) $\hat{\phi}_3(s)$ with values in the mOFC, (b) $\hat{\phi}_4(s)$ with values in the aINS, (c) $\hat{\phi}_{13}(s)$ with values in the parietal cortex, (d) $\hat{\phi}_5(s)$ with values in the mOFC, VLPFC and DLPFC, (e) $\hat{\phi}_9(s)$ with values in the mOFC and the parietal cortex.

Risk Score

For each task, the response was either a binary decision where the subject rates the stream (risky or safe) or it was a numerical rating of either the expected return or the perceived risk rating of the investment itself. We convert the responses into a binary result where 0 means the subject rated the investment return stream as low risk and 1 means they rated it high risk. This binary is determined directly for tasks 3 and 23 as they have binary outcomes. For tasks 1 and 26 we have divided the responses by the median. The risk prevalence is each task is summarised in Table (4.6). Denote the resulting vector as R, comprised of individual risk responses R_{iq}. The scores estimated from the PCs containing the ROIs are shown in Fig.(4.7). Define $\vartheta_{il} = Var(\hat{\psi}_{il}(t))$ for $l \in \{1, \ldots, 14\}$ to be the total variation over time for a score function. We assume that the tasks have different risk prevalence by adding a coefficient stratifying based on tasks but we assume the ROIs have the same effect size across task types. We set the general linear model (GLM) with

	Task 1	Task 3	Task 23	Task 26
Risk prevalence	0.4	0.8	0.93	0.6

Table 4.6: Percentage of subject responses which classified the investment as risky.

the logit link function:

$$\Pr(R_i = 1) = \mathrm{logit}(\beta_0 + \sum_{l \in L} \beta_l \vartheta_{il}). \tag{4.19}$$

At first L is set to 14 and the model is estimated to determine significant covariates. These are found to be ϑ_{il} for $l \in \{3, 4, 5, 9, 13\}$ and they are used to estimate the final logistic regression model as equation (4.19) where $l \in \{3, 4, 5, 9, 13\}$ whose estimates are summarised in Table (4.7). Overall, the most significant score functions were ϑ_{i4} and ϑ_{i5} which correspond to $\hat{\phi}_4(s)$ and $\hat{\phi}_9(s)$ which correspond the activity of aINS, mOFC and the parietal cortex.

	β_1	β_2	β_3	β_4	β_5
l	3	4	5	9	13
Coef.	-1.05	5.61	-0.43	-7.29	0.64
Pr	0.66	0.029	0.81	0.05	0.63

Table 4.7: GLM coefficients indexed by their order in the regression, the ℓ that indicates which PC correspond to each coefficient and the p-values.

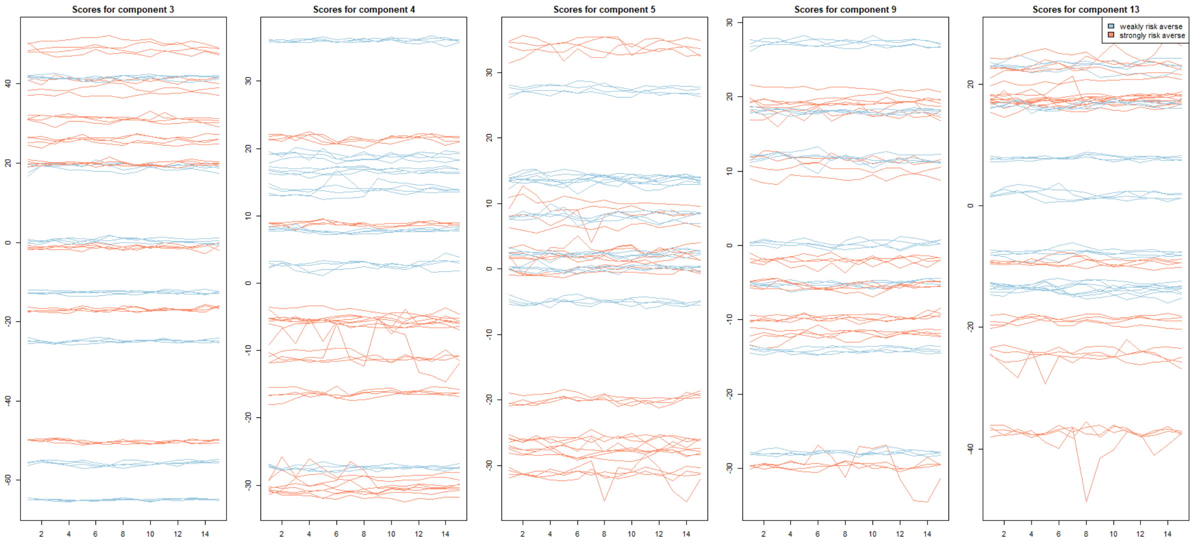


Figure 4.7: Score functions corresponding to the PCS from Fig 4.6. Blue represents weakly averse and orange represents strongly risk averse subjects.

.

Regions of Interest

The locations where $\hat{\phi}(s)$ have nonzero values are marked as red area in Fig. (4.6) which presents the estimated PCs which have highlighted ROIs found previously in the literature (superimposed over a mean subject image). Given the defined brain area we are able to analyse, we are able to recover all 5 regions associated with risk in the literature. These are: mOFC, aINS, VLPFC, DLPFC and the parietal cortex.

	β_1	β_2	β_3	β_4	β_5
l	3	4	5	10	12
Coef.	-0.049	-0.02	-0.07	-0.09	0.10
Pr	0.12	0.48	0.21	0.06	0.03

Table 4.8: GLM coefficients obtained for Li et al., 2019 by their order in the regression, the ℓ that indicates which PC correspond to each coefficient and the p-values.

Method by Li et al., 2019

To compare our model with previous methods, the same dataset was analysed using the approach introduced by Li et al., 2019. Their method did not directly consider the time component but instead took the difference of the first 3 images to account for brain activity at the beginning of a task. The code used for this comparison was sourced from the github link available in their paper. The estimated FPCA model for this dataset was able to recover at most 66.6 % VE with 30 PCs. The ROIs were extracted from the estimated PCs by taking the upper and lower quantiles. These new trimmed PCs representing the binary ROIs are used to estimate new scores υ_{il} which can be used in a logistic regression as in equation (4.19), to associate regions with risk response. Covariates that in univariate analysis were found to be significant were selected to be υ_{il} for $l \in \{3, 4, 5, 10, 12\}$ and their corresponding identified regions are shown in Figure (4.8). The selected scores corresponding to the trimmed PCs were used in a GLM similar to eq. (4.19) where the outcome variable is a binary risk score and another variable accounting for task index was included. the results of this regression are shown in Table (4.8) which have found PCs 10 and 12 to be most significant. These PCs correspond mOFC, VLPFC and the parietal cortex. Two out of the three identified regions were found using our approach, namely mOFC and the parietal cortex. Both methods were able to recover the same key regions in this dataset, and were able to identify two of the same significant regions. The difference being that our approach found aINS to be significant, whilst Li et al., 2019 found VLPFC. Notably, the PCs and the ROIs as a result of the trimmed PCs are different, as shown in Figures (4.8) and (4.6).

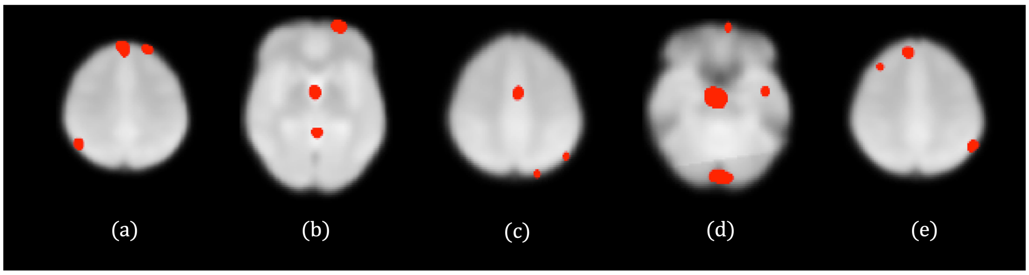


Figure 4.8: ROIs recovered from the application of Li et al., 2019 (a) $\hat{\phi}_3(s)$ with values in the parietal cortex and VLPFC, (b) $\hat{\phi}_4(s)$ with values in the mOFC (c) $\hat{\phi}_5(s)$ with values in the VLPFC and parietal cortex, (d) $\hat{\phi}_{10}(s)$ with values in the mOFC and VLPFC, (e) $\hat{\phi}_{12}(s)$ with values in the parietal cortex.

4.5.4 Data Analysis Conclusion

Overall, we were able to reconstruct the data with VE 90.1% given $L = 14$ estimated components. This is one less than the number of individuals which suggests that interpersonal variability is higher than that across tasks, as adding different tasks from the same subject did not increase the available degrees of freedom.

Logistic regression was able to identify ROIs and associate them with risk propensity of the subjects. We were able to recover all available regions in the data. These were aINS, mOFC, VLPFC, DLPFC and the parietal cortex. Finally we created a model for the association of said region to risk, and found aINS, mOFC and the parietal cortex to be most significant is determining the risk preferences of subjects.

Comparison to Li et al., 2019 showed that our method outperformed theirs in terms of VE, with their being able to explain 66.6% of data variation with more components, but both were able to recover the same ROIs. Both methods found mOFC and the parietal cortex to be significant regions contributing to a subject's risk response.

4.6 Conclusion and Discussion

We introduced a spatio-temporal FPCA model applicable for any dense temporal brain imaging stored in arrays. The model represents the data as a linear product of two functions, one defined over space and the other defined over time. This approach preserved the spatial structure and the relationships between areas within the array and can extract important features using a non-parametric dimensionality reduction. Each of the extracted elements can then be modelled for activity over time which has a particular use case during fMRI studies.

This model is estimated with a fast method that circumvents the computational burden of estimating a covariance matrix. We have designed a simulation study to evaluate the estimation method in identifying the true number of components, ability to recover true underlying functions and reconstructing the data.

In cases when the underlying model matched the fitted one, our method was able to correctly estimate the score functions and PCs whilst also identifying the appropriate number of components. The estimation accuracy would be lowered when more noise was present and when the underlying functions were more complex themselves, however, it remained above 90% for all but one simulation. In simulations when the underlying model deviated from the fitted one, we find that the overall accuracy of the reconstruction decreases, and functions that specifically deviate from model assumptions have worse estimations. Throughout the simulation, we found that the choice of type and number of basis vectors for the estimation of the score functions plays an important role in the accuracy of the reconstructions.

For the data analysis, our model was able to reconstruct the data with VE 90.1% with 14 estimated components. This was consistent with our findings from the simulation study and outperformed Li et al., 2019 on the same data in terms of VE. Considering results presented in previous papers that do not model time dependency, we got a similar result to Chen et al., 2014 got 94% with 20 PCs and Li et al., 2019 got 80% with 9 PCs. We were able to identify all ROIs (aINS, mOFC, VLPFC, DLPFC and the parietal cortex) and use the in a logistic regression to determine associations between risk-response and brain activity in POIs. And with logistic regression we were able to find the strongest associations to be in the aINS, mOFC and parietal cortex.

Comparing our method to the one introduced by Li et al., 2019 shows that we are able to achieve a higher VE on a smaller dataset. Our inclusion criterion was different as we sought to analyse the data that provided additional information and hence, we have only considered 4 tasks. In their paper, the authors included all 81 tasks, which could potentially be the reason for the discrepancy between the results in their paper and the ones presented here. Their approach introduced more complexity via taking the difference of the first three images which could explain how they benefited from a larger sample size. Our approach doesn't require input regarding the temporal element and hence the implementation is more straightforward but also resulted in different results, as we have seen repeated behaviour patterns within subjects when doing the same task. Both methods were able to recover the same ROIs however the PCs estimated were different.

Our analysis found that interpersonal variability plays a strong role in analysing behaviour across tasks as each subject followed a specific pattern of behavior that limited the number of tasks that were used for the analysis. This is different from methods that concatenate multiple time points into one image, where this effect is lost. Our findings withing the fMRI data matched the findings of previous analysis approaches in terms of region identification. We approached the analysis differently on an individual task basis rather than the whole study approach, as in our preliminary data analysis subject behaviour was repeatable over time.

Our model and estimation method shows a new approach in analysing temporal data using a non-parametric approach, suitable for use on fMRI data for association analysis. We have not only shown that it is computationally efficient, but it outperforms existing methods on small data and uncovers different trends that previous analysis. Studying the behaviour of the estimated functions would help better understand the significance of each of the estimated components. This model still lacks some of the usual elements such as significant and confidence intervals which would help in the task of identifying ROIs and understanding which elements in the estimated functions show important information. The estimation method could be further improved by incorporating different regression models. In particular, equation (4.14) assumed for each voxel in the raw estimate of the PC to be independent prior to being smoothed, one could relax this assumption with a different regression model.

In the chapter we have compared our method to ones available for image analysis, however models that decompose functions over time and space are available in lower dimensions. In the next chapter, we will investigate the performance of our model in lower dimensions and compare them to existing methods to better understand the benefits and limitations of our approach.

Chapter 5

Studying STFPCA Model Performance in Low Dimensions

This chapter will describe three Functional PCA models and compare them in a simulation study. We will look at their performance in estimation in various settings, computation time and other factors such as the number of basis functions used in the estimation procedure as well as sampling density.

5.1 Introduction

In Functional Principal Component Analysis, a random function $X(s,t)$ is decomposed into a sum of random variables multiplied with a set of optimal basis functions that each maximise the variation of $X(s,t)$. This decomposition is achieved by applying the Karhunen-Loève Theorem and takes the form:

$$X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty} \psi_j v_j(s,t), \tag{5.1}$$

where $v_j(t)$ are the optimal basis functions, equal to the normalized eigenfunctions of the sample covariance operator, and ψ_j are the principal component scores equal to $\langle X(t), v_j(t) \rangle$.

FPCA can take different forms, however we are interested in decompositions that separates the domains s and t. This category of models can be broadly represented using the equation:

$$X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty} \psi_j(t)u_j(s), \tag{5.2}$$

where $\psi_j(t)$ would now be the score functions varying over time and $u_j(s)$ would be the principal components whose definition can vary.

Several models have been developed to approach this problem in the context of FDA. Greven et al., 2010 proposed a model where the variation of functional data is decomposed into a baseline subject-specific variability, longitudinal subject-specific variability, subject-visit-specific variability. Their proposed model can be viewed as the functional extension of a longitudinal mixed effects model where random effects are replaced by random processes. The time component is incorporated into the model through a linear structure which relies on additive assumptions. In equation (5.2) it would take the form $\psi_j(t) = \zeta_{0jr} + t\zeta_{1jr}$ where ζ_{0jr}, ζ_{1jr} are random terms. This linear structure can be limiting when working on imaging and hence we do not consider it in this simulation.

Chen and Müller, 2012 propose a model without the additive assumptions of Greven et al., 2010. The score functions take the form $\psi_j(t) = \sum_{k\geq 1} \zeta_{jk}\kappa_{jk}(t)$ with $\kappa_{jk}(t)$ orthogonal basis functions and a random coefficient ζ_{jk}. This is achieved by a two-step Karhunen-Loève expansion. However, their model uses time varying basis functions $u_j(s|t)$ and proved to be an order of magnitude more computationally expensive and less accurate compared to the model proposed by Park and Staicu, 2015, who propose a model we will discuss in this chapter.

Park and Staicu, 2015 also applied a two step KL expansion for the scores without assuming a parametric structure on the covariance matrices. Hence the score function takes the same form as before, namely $\psi_j(t) = \sum_{k\geq 1} \zeta_{jk}\kappa_{jk}(t)$. Their model assumes a non-time dependant PC $u_j(s)$. In contrast to the two previous models and ours, they assume the residual process to be a sum of a random square integrable function at a discrete time point t, $\varepsilon_{1t}(s)$ with a covariance $Cov_t(\varepsilon_{1t}(s), \varepsilon_{1t}(s'))$ and white noise denoted $\varepsilon_{2t}(s)$.

Our proposed model imposes no parametric assumptions on the score functions and therefore is comparable to the latter two models. In contrast to Park and Staicu, 2015 who derive their PCs from the marginal covariance of $X(s,t)$, we derive the PCs from covariance of the marginal process. As the model proposed Park and Staicu, 2015 have outperformed Chen et al., 2015 in terms of accuracy and computation times, our simulation will be comparing our model to FPCA and Park and Staicu, 2015.

We will further define the marginal process model from Chapter 4 and introduce the model proposed by Park and Staicu, 2015. We will prove the theoretical properties of the estimates and compare the models in a simulation study.

5.2 Model Description

The following section summarises the methods introduced in Chapter 4 and then describes the model proposed by Park and Staicu, 2015.

5.2.1 Marginal Process Model

Consider a random function $X(s,t) \in L^2$ defined on the bounded intervals S and T, with $s \in S, t \in T$. We assume $X(s,t)$ to be square integrable and hence it has a unique mean functions $\mu(s,t) = \mathbb{E}[X(s,t)]$. We can decompose X into a sum of the mean and the variation $U(s,t)$:

$$X(s,t) = \mu(s,t) + U(s,t), \tag{5.3}$$

with $U(s,t) \in L^2$ a random function with $\mathbb{E}[U(s,t)] = 0$

To model $U(s,t)$ Consider the marginal function $U(s) = \int U(s,t)g(t)dt$ where $g(t)$ is the sampling density function over t, $g(t)$ is continuous and $sup_{t\in\mathcal{T}}(g(t)) < \infty$. The function $U(s)$ has a covariance $\nu_t(s,s') = \mathbb{E}\Big[U(s)U(s')\Big]$ which is the kernel to the covariance operator $\varsigma_t(U)(s)$ defined in eq. (4.2). The covariance function ν, the kernel of the covariance operator can be expressed as the eigenfunctions $\phi_j(s)$ of the covariance operator (eq. (4.3)) and so

$$\nu(s,s') = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(s'). \tag{5.4}$$

The eigenfunctions form a time-invariant orthonormal basis in $L^2(S)$ and are different from those in Park and Staicu, 2015 as they are optimized minimize the function $S(\theta(\cdot))$

with vector $\theta(\cdot) = ((\theta_1(\cdot), \dots, \theta_K(\cdot))^T$ containing K arbitrary basis functions $\theta(\cdot)$:

$$S_p(\theta(\cdot)) = \int \mathbb{E}\big|\big|U_i(\cdot, t) - \sum_{k=1}^{K} \langle U_i(\cdot, t), \theta_k(\cdot)\rangle \theta_k(\cdot)\big|\big|^2 g^2(t)dt.$$

By applying Mercer's theorem and the KL theorem, the process $U(s)$ can be expressed as an infinite linear combination of the deterministic eigenfunctions $\phi_l(s)$ with random uncorrelated weights $\omega_l = \langle U, \phi_l\rangle$:

$$U(s) = \sum_{l=1}^{\infty} \omega_l \phi_l(s). \tag{5.5}$$

We propose this new $\phi_l(s)$ to be the basis function of new decomposition of $U(s,t)$ together with the random score functions $\psi_l(t) = \langle U(s,t), \phi_l(s)\rangle$. They also have a covariance function denoted as $G_l(t,t') = Cov(\psi_l(t), \psi_l(t'))$ which is a smooth function defined on $T \times T$. Using Mercer's theorem, it can be expressed as in eq. (4.5) with where $\kappa_{k1} \geq \kappa_{k2} \geq \cdots \geq 0$ and $\{\xi_{lm}(t)\}$ forming an orthonormal basis in L^2. Using the Karhunen-Loève theorem, the $\psi_l(t)$ can be expressed using this basis $\{\xi_{lm}(t)\}$ as in eq. (4.6). Finally, we can represent $U(s,t)$ as:

$$U(s,t) = \sum_{l=1}^{\infty} \psi_l(t)\phi_l(s) = \sum_{l=1}^{\infty}\sum_{m=1}^{\infty} \eta_{lm}\xi_{lm}(t)\phi_l(s) \tag{5.6}$$

Using the above decomposition, we define the population model:

$$X(s,t) = \mu(s,t) + U(s,t); \quad U(s,t) = \sum_{l=1}^{\infty} \psi_l(t)\phi_l(s) \tag{5.7}$$

where $\psi_l(t) = \sum_{m=1}^{\infty} \eta_{lm}\xi_{lm}(t)$ are the score functions and $\phi_l(s)$ are the time-invariant PCs. If one chooses L and M PCs and basis functions of the scores to sufficiently reconstruct $U(s,t)$ the result is the truncated model:

$$U(s,t) = \sum_{l=1}^{L} \psi_l(t)\phi_l(s) = \sum_{l=1}^{L}\sum_{m=1}^{M} \eta_{lm}\xi_{lm}(t)\phi_l(s). \tag{5.8}$$

5.2.2 Marginal Covariance Model

We now introduce the model from Park and Staicu, 2015. Their paper specifically deals with longitudinal observations however we will generalise their approach to a dense t. Define $X(s,t) \in L^2$ as before and consider its variation $U(s,t) \in L^2$:

$$X(s,t) = \mu(s,t) + U(s,t) + \varepsilon_1(s,t) + \varepsilon_2(s,t), \tag{5.9}$$

with $\varepsilon_1(s,t)$, denoting a random function of variation not covered in $U(s,t)$ with a covariance function $\Gamma_t(s,s') = Cov(\varepsilon_1(s,t), \epsilon_{1i}(s',t))$. White noise is denoted with $\varepsilon_2(s,t)$ and $Cov(\varepsilon_2(s,t), \epsilon_2(s',t)) = \sigma^2$ for $s = s'$ and 0 otherwise. The covariance function of $U(s,t)$ is $c((s,t),(s',t')) = \mathbb{E}[U(s,t)U(s',t')]$ and the marginal covariance

$$\varsigma(s,s') = \int c((s,t),(s',t))g(t)dt, \tag{5.10}$$

where $g(t)$ is the sampling density of t that is continuous and has an upper bound. If we consider the bivariate process $W(s,t) = U(s,t) + \varepsilon_1(s,t)$, this has the marginal covariance of the form

$$\Xi(s,s') = \varsigma(s,s') + \Gamma_t(s,s'). \tag{5.11}$$

Consider the eigendecomposition problem:

$$\int_{\S} \Xi\big((s),(s')\big)\varphi_k(s)dtds = \lambda_k\varphi_k(s'), \tag{5.12}$$

where $\{\varphi_k(s_r)\}$ are the eigenfunctions that form a time-invariant orthonormal basis in L^2 positive and ordered eigenvalues λ. These functions optimise for the following mean squared error:

$$S_c(\theta(\cdot)) = \int \mathbb{E}\big|\big|U_i(\cdot,t) - \sum_{k=1}^{K} < U_i(\cdot,t), \theta_k(\cdot) > \theta_k(\cdot)\big|\big|^2 g(t)dt \tag{5.13}$$

Applying the Karhunen-Loève Theorem yields:

$$X(s,t) = \mu(s,t) + \sum_{k=1}^{\infty} \psi_k(t)\phi_k(s) \tag{5.14}$$

where $\varphi_k(s)$ is a basis $\in L^2$ from equation (5.12) and $\psi_k(t)$ are the corresponding coefficients that are zero mean random functions that can be correlated over t with a smooth covariance function $G_k(t,t') = Cov(\psi_k(t), \psi_k(t'))$. Then by Mercer's Theorem

$$G_k(t,t') = \sum_{p\geq 1} \kappa_{kp}\xi_{kp}(t)\xi_{kp}(t') \tag{5.15}$$

where $\kappa_{k1} \geq \kappa_{k2} \geq \cdots \geq 0$ and $\{\xi_{kp}(t)\}$ form an orthonormal basis in L^2. Using the Karhunen-Loève theorem we get the expression:

$$\psi_{ik}(t) = \sum_{p=1}^{\infty} \eta_{ikp}\xi_{kp}(t) \tag{5.16}$$

where $\eta_{ikp} = \int \psi_{ik}(t)\xi_{kp}(t)dt$ are random variables uncorrelated over p with zero mean and variance equal to κ_{kp}. This is the final population model but can be written as equation (5.14)

$$U(s,t) = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \eta_{ikp}\xi_{kp}(t)\varphi_k(s). \tag{5.17}$$

Similarly to before, it is fair to assume that K and P of the PCs and score components would explain a sufficient amount of the data, so this model as well has the truncated form:

$$U(s,t) = \sum_{k=1}^{K}\sum_{p=1}^{P} \eta_{ikp}\xi_{kp}(t)\varphi_k(s). \tag{5.18}$$

5.3 Simulation

We have introduced three models that provide a way of representing a function defined over two domains. These are the classic FPCA approach as in equation (5.1), the marginal process model from equation (5.6) and the marginal covariance model from equation (5.16). In the following simulations we will be comparing the truncated FPCA model of the form $X(s,t) = \mu(s,t) + \sum_{j=1}^{J} \psi_j v_j(s,t)$ with the truncated models from equations (5.8) and (5.18) to refer to the marginal process model and the marginal covariance model, respectively. We are interested in how these models perform on in different conditions, so different simulation designs will reflect data where the underlying model is a product of two functions set over two separate domains and ones where the underlying models include functions defined over two domains with a non-linear relationship. Design 1 will reflect the former case case, whereas Design 2 will reflect the latter.

We will primarily focus on the marginal process and covariance models to evaluate their reconstruction ability and computation time. This will be measured using Variance Explained (VE) and integrated squared error (ISE). Furthermore, we will investigate the effect that the choice of basis vectors and sampling density has on the model estimation.

5.3.1 Designs

The data will be generated on $s, t \in [0, 1]$ using the following functions:

$$\begin{aligned} \xi_{1i}(t) &= a_i \cos(5.5t) & \xi_{2i}(t) &= b_i \sin(5t) & \xi_{3i}(t) &= \cos(d_i t) \\ \omega_1(s) &= \cos(6\pi \cdot (s)) & \omega_2(s) &= \sin(4\pi \cdot (s)) & \omega_3(s,t) &= c_i \sin(6\pi \cdot (st)) \end{aligned}$$

where $a_i \sim N(1.5, 1.25), b_i \sim N(2, 1.5), c_i \sim N(1, 1.3), d_i \sim N(2, 3)$.

Design 1

Both designs generate data that would fit the assumptions imposed by our models. Design 1.1 will be a simpler design where the random variable is outside the trigonometric function, whereas Design 1.2 will utilise $\xi_{3i}(t)$ which has the random element within the function. The data will be generated as follows

$$\begin{aligned} &\text{Design 1.1: } X_i(s,t) = \xi_{1i}(t) \cdot \omega_1(s) + \xi_{2i}(t) \cdot \omega_2(s) + \varepsilon_i(s,t) \\ &\text{Design 1.2: } X_i(s,t) = e_i \cdot \cos(s) \cdot \xi_{3i}(t) + \varepsilon_i(s,t), \end{aligned}$$

with $e_i \sim N(\sqrt{3}, 0.5)$ and where $\varepsilon_i(s,t)$ is idiosyncratic noise that will vary between settings.

Design 2

Design 2.1 follows Design 1.1 but includes an extra term that changes the way the data is generated. Design 2.1 has a dependency on s and t together and Design 2.2 includes a random term in the function defined over s which means that there is an unknown number of PCs. The data will be generated as follows

$$\text{Design 2.1: } X_i(s,t) = \xi_{1i}(t)\cdot\omega_1(s) + \xi_{2i}(t)\cdot\omega_2(s) + \xi_{3i}(t)\cdot\omega_3(s) + \varepsilon_i(s,t)$$
$$\text{Design 2.2: } X_i(s,t) = e_i\cdot\cos(f_i s)\cdot\xi_{3i}(t) + \varepsilon_i(s,t),$$

with $e_i \sim N(\sqrt{3}, 0.5)$ and $f_i \sim N(1.5, 0.5)$ where $\varepsilon_i(s,t)$ is idiosyncratic noise.

Evaluation

Our simulation will primarily focus on designs 1.1 and 2.1, and hence these simulations will be run with more settings. Designs 1.2 and 2.2 are used for additional information but will not be included in additional experiments described below.

For 1.1 and 2.1, we will vary the number of subjects $n \in 50, 100, 1000$ and the noise ε_{ijk} that can be one of three settings:

$$\text{1: } \varepsilon_{ijk} = 0,$$
$$\text{2: } \varepsilon_{ijk} = N(0,\ 0.16^2),$$
$$\text{2: } \varepsilon_{ijk} = N(0,\ 0.32^2).$$

The functions are sampled at 60 equidistant points. Designs 1.2 and 2.2 will be run with $n = 1000$ and the noise will be either 0 or sampled from one of the distributions $N(0, 0.1)$ and $N(0, 0.2)$. Here these functions are sampled over equidistant 30 points.

Prior to the estimation of the model a mean function is estimated and removed from the data and the remaining function will be denoted as $U_i(s,t)$. The score functions for the marginal process and marginal covariance models will be estimated using 11 b-spline vectors as those have shown to be most effective in individual testing, yielding the highest VE per basis vector.

We will use Designs 1.1 and 2.1 to further understand the model, looking specifically at the effect of sampling density and the number of basis function representing $\xi(t)$.

To understand the effect of the number of b-spline vectors on the reconstruction of the scores, we will conduct a separate simulation where we vary the number of vectors to see how many are necessary for complete reconstruction. As the s domain is dense the estimation of principal components is done without the use of basis vectors and hence FPCA does not have this dependency. If a score function can be represented with vectors as follows: $\xi(t) = \sum_{p=1}^{P} c_p b_p(t)$ with $b_p(t)$ denoting b-spline vectors, we will vary P from 3 to 15. In this simulation, we set $n = 100$ and $\varepsilon_i(s,t) \sim N(0, 0.16)$ and run the estimation procedure as normal. We will look at how the value of b affects the VE and what would be the optimal choice of b meaning that any additional vector would not increase VE significantly.

Finally, would like to understand how the sampling of curves can affect the estimation of the model functions. Here we would create 60 sampling points along s that would be considered the full observed function, we will then 'observe' either 30, 20 or 10 points along s at equidistant points. So if the complete function $\omega_1(s)$ is sampled at s_j for $j \in \{1, \ldots, 60\}$ and $s_j \in \{\frac{1}{60}, \ldots 1\}$. Then we will run 3 simulations where $j \in \{1, \ldots, 30\}, \{1, \ldots, 20\}$ and $\{1, \ldots, 10\}$. This simulation will only run for D2.1 with $n = 100$.

Given a model with L computed PCs, it will be evaluated using VE:

$$\mathrm{VE}(L) = 1 - \left[\frac{\sum_{i=1}^{n} \int\int \left[U_i(s,t) - \sum_{l=1}^{L} \psi_{il}(t)\phi_l(s)\right]^2 dsdt}{\sum_{i=1}^{n} \int\int \left[U_i(s,t)\right]^2 dsdt}\right]. \tag{5.19}$$

We will also investigate the effect that the number of basis vectors chosen in the estimation of a score function will have on the reconstruction. This will be measured using VE.

To understand how sampling a function over s affects the estimation procedure, we will use VE and integrated root squared error:

$$\mathrm{IRSE}(\hat{\phi}(s)) = \int \sqrt{[\phi(s) - \hat{\phi}(s)]^2} ds \tag{5.20}$$

We will evaluate the estimated PC functions in two ways: first, we will directly compute the difference between the true and estimated function at the points we sample, second, we will smooth the PC function and compute the full difference between the true and the smoothed estimated function.

The simulations will run on the ARC4 computer with Intel Xeon Gold 6138 CPUs using their standard nodes and run time will be recorded and compared between models. For each combination there will be 100 replicates. For each generated sample, we will estimate three models: FPCA as introduced in the Chapter 3, the Marginal Covariance model from Park and Staicu, 2015 and our introduced Marginal Process model.

Error	$VE1_p$	$VE1_c$
0	0.9419 (0.002)	0.9419 (0.002)
10%	0.9175 (0.004)	0.9175 (0.004)
20%	0.8779 (0.005)	0.8779 (0.005)

Table 5.1: D1.2 result: VE. The values in the table represent the mean and standard deviation (in brackets) over 100 replicates. subscripts p and c refer to results obtained from the marginal process and the marginal covariance models.

5.3.2 Results

We will now discuss the results from the simulation designs, subscripts p, c and f refer to estimated functions or results obtained from the marginal process model, the marginal covariance model and FPCA, respectively. We will discuss the reconstruction ability, influence of the number of basis vectors and the density of sampling points.

Reconstruction Ability

We will first consider the cases of Designs 1.1 and 1.2, where the underlying models match the fitted model. Table (5.2) has the variance explained for design 1.1 showing the number of estimated components and the VE for each model. The standard deviations can be found in the appendix (Table (A.1)). Overall, looking at Table (5.2), the marginal process and covariance models have a similar cumulative VE with 2 principal components, with the noise levels not changing their estimation largely. Both models, however, have a slightly lower VE (0.997) than FPCA (1). It appears that the average VE over 100 replicates is not much altered by an increasing sample size (0.781 for $n = 50$ versus 0.786 for $n = 1000$). However, the sample size does affect the standard deviation, with it becoming smaller as n increases (Table (A.1)).

Design 1.2 includes the random element within the trigonometric function which can increase estimating difficulty for the score function, however it only had 1 PC to estimate. The resulting VE from the models is shown in Table (5.1). In the case of design 1.2, both

Noise	Nr of PCs			Variance Explained					
	Proc	Cov	FPCA	Proc		Cov		FPCA	
n=50				$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$
0%	2	2	2	0.555	0.997	0.566	0.997	0.573	1
10%	2	2	2	0.510	0.909	0.520	0.910	0.529	0.913
20%	2	2	2	0.443	0.781	0.452	0.783	0.462	0.788
n=100									
0%	2	2	2	0.557	0.997	0.563	0.997	0.569	1
10%	2	2	2	0.512	0.911	0.518	0.911	0.524	0.914
20%	2	2	2	0.444	0.783	0.450	0.784	0.456	0.787
n=1000									
0%	2	2	2	0.566	0.997	0.566	0.997	0.569	1
10%	2	2	4	0.521	0.912	0.522	0.912	0.524	0.914
20%	2	2	20	0.454	0.786	0.455	0.786	0.456	0.787

Table 5.2: D1.1 result: average VE per PC (cumulative) over 100 replicates.

Error	$\hat{\psi}_p(t) - \hat{\psi}_c(t)$	$\hat{\phi}_p(s) - \hat{\phi}_c(s)$
0	0.2270 (2.898)	0.0070 (0.087)
10%	-0.1275 (3.584)	-0.0035 (0.107)
20%	-0.1256 (3.603)	-0.0035 (0.106)

Table 5.3: D1.2 result: MISE between scores and PCs from models (5.8) and (5.18). Values in the table represent the mean and standard deviation (in the brackets) over 100 replicates.

models have the same VE up to 4 significant figures. To confirm that the estimated functions are indeed different, the difference was taken and averaged over 100 replicates shown in Table 5.3. Comparing this to the results from Design 1.1, we can see that for the zero noise scenario, the VE was lower in D1.2 (0.942) than for D1.1 (0.997).

When looking at D2.1, let us look at Tables (5.5), (5.6) that show the number of components estimated and the resulting VE, respectively. The standard deviations can be seen in the appendix as Table (A.2). Both FPCA and the marginal process models are able to identify 3 PCs whereas the covariance model is now estimating between 4 to 5 PCs. Considering the VE, we can see that FPCA is able to reconstruct data fully without much difference across the sample size and noise level. The marginal covariance model achieves lower VE than FPCA with more components but is still able to explain a sufficient amount of variation (0.98 with no noise and 0.87 at the highest noise setting), notably, this value is only slightly lower than for D1.1 where the model achieved 0.997 VE with no noise present. Notably, the first and second estimated components correspond to $\omega_1(s)$ and $\omega_2(s)$, and any following estimated functions make up $\omega_3(st)$. Finally, con-

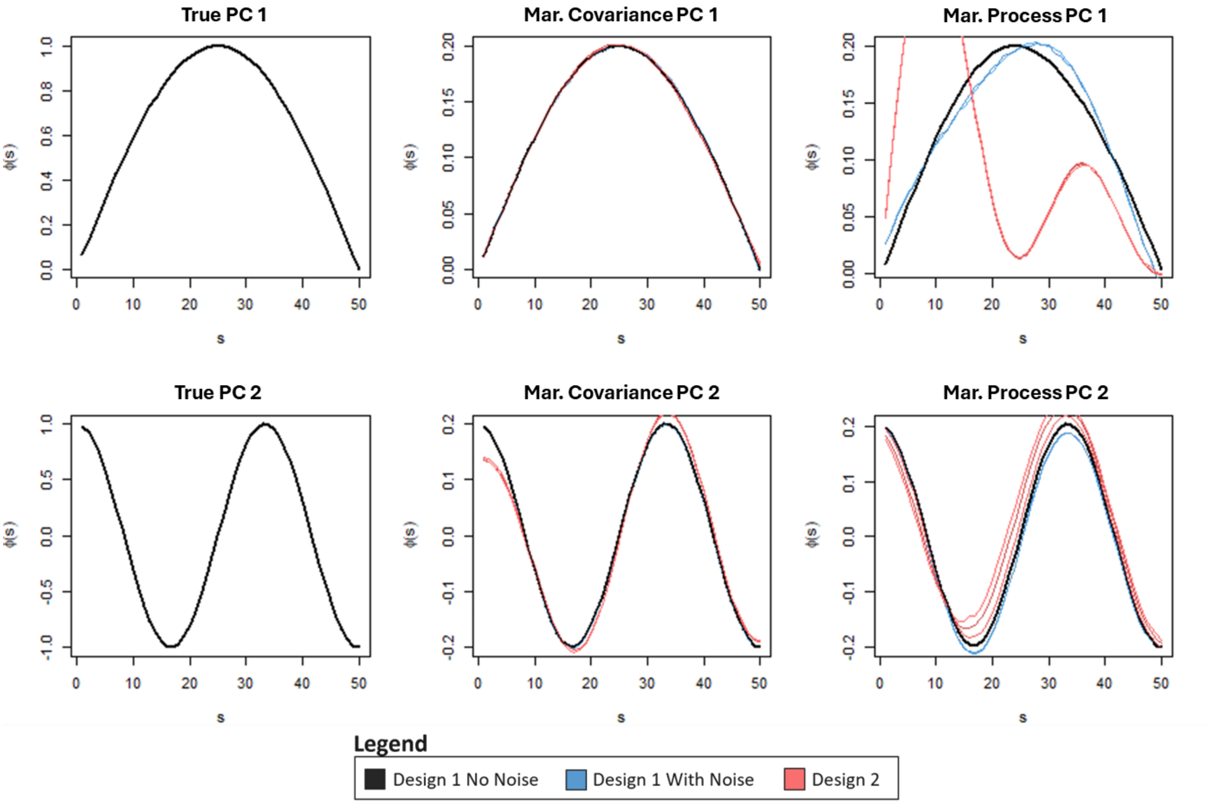


Figure 5.1: Plots of the true and estimated PCs from both models with n=1000. The black line is estimated with D2.1 and no noise, the blue lines are the other noise settings for D2.1. Orange lines correspond to all D2.2 noise settings.

sidering the marginal process model we can see that it has the lowest cumulative VE and can at best recover 0.87 of the variation compared to the covariance model that achieves 0.98. However, considering the performance with 3 PCs, both models have similar VE (the marginal process model achieved between 0.77-0.87 and marginal covariance model achieved 0.78-0.87).

The standard deviation was computed based on the cumulative VE up to some PC. The general trend for the standard deviation (Table (A.2)) is that it goes down as the sample size goes up. The standard deviation appears to be lower when the PC corresponds to less VE. In practice, this means that as the PC index increases, the sd tends to go down. This is very clearly shown for the marginal covariance and FPCA models. The only deviation from this trend is the marginal process model for $\hat{\phi}_2(s)$, as the standard deviation there is higher than for $\hat{\phi}_1(s)$ (for example for $n = 50$ with no noise sd for $\hat{\phi}_2(s)$ is 0.0308 and for $\hat{\phi}_1(s)$ it is 0.0185). This could be explained by the fact that the second PC explains more VE than the first, suggesting a change in PC order, which is seen in Figure (5.2).

Error	$VE1_p$	$VE1_c$	$VE3_p$	$VE3_c$
0	0.6021 (0.016)	0.5881 (0.015)	0.9416 (0.002)	0.9417 (0.002)
10%	0.5844 (0.016)	0.5711 (0.015)	0.9129 (0.002)	0.9129 (0.002)
20%	0.5574 (0.014)	0.5447 (0.014)	0.8658 (0.005)	0.8658 (0.005)

Table 5.4: D2.2 result: VE. The values in the table represent the mean and standard deviation (in brackets) over 100 replicates.

Figure (5.1) shows the estimated PCs for all the settings of the simulation. The marginal covariance model displays less variation subject to the model fit and the noise, with both PCs being able to recover their original shape relatively well. The marginal process can recover the true shapes of the PCs but with slight errors subject to noise, one can see that the blue line that depicts the estimates for D1 recovers the original shape but is slightly skewed. However, in D2 the first PC is not matching the design, whilst PC 2 is of appropriate shape but appears to have a slightly wider range.

Noise	**Variance Explained**										
	Proc			**Cov**					**FPCA**		
n=50	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_3(s)$	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_3(s)$	$\hat{\phi}_4(s)$	$\hat{\phi}_5(s)$	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_3(s)$
0%	0.260	0.580	0.868	0.439	0.760	0.871	0.973	0.982	0.476	0.789	1
10%	0.251	0.560	0.836	0.424	0.732	0.840	0.938	0.947	0.461	0.763	0.965
20%	0.232	0.515	0.768	0.392	0.673	0.773	0.863	0.873	0.430	0.706	0.890
n=100											
0%	0.257	0.581	0.868	0.435	0.760	0.872	0.973	0.982	0.465	0.776	1
10%	0.247	0.559	0.836	0.420	0.732	0.839	0.937	0.947	0.451	0.749	0.963
20%	0.228	0.514	0.768	0.388	0.673	0.772	0.862	0.871	0.420	0.694	0.887
n=1000											
0%	0.258	0.581	0.868	0.437	0.759	0.871	0.973	0.982	0.463	0.777	1
10%	0.249	0.559	0.836	0.422	0.731	0.839	0.938	0.947	0.449	0.751	0.963
20%	0.229	0.514	0.770	0.390	0.673	0.773	0.864	0.873	0.419	0.695	0.887

Table 5.5: D2.1 result: average VE per PC (cumulative) over 100 replicates.

Finally for D2.2, where the function over s has a random component in it, we have Table (5.4) that shows the VE for both models. Both methods are able to estimate the same number of PCs (3) and achieve a similar VE with 3 PCs (0.94). The first PC seem to explain a slightly different amount of VE (0.6 for marginal process versus 0.59 for marginal covariance). This difference can be seen in Figure (5.2) with PC1 being different curves, however, PC 2 and PC 3 appear to be similar. Both models have a similar VE and MSE across the 3 noise settings, and comparing the results from D1.2 and D2.2, we can see that in the simulation with no noise, the reconstruction was similar. Comparing Tables

(5.1) and (5.4), both designs have a similar reconstruction (0.942 in D1.2 and 0.942 in D2.2 for both models up to three significant figures).

Noise	**Nr of PCs**		
n=50	**Process**	**Covariance**	**FPCA**
0%	3	4.63	3
10%	3	4.72	3
20%	3	4.81	3
n=100			
0%	3	4.57	3
10%	3	4.64	3
20%	3	4.85	3
n=1000			
0%	3	4.98	3
10%	3	4.99	3
20%	3	5	12

Table 5.6: Design 2: Average number of estimated components for each model.

Table (5.7) shows the estimation times between the two models in seconds. The time measured for the marginal process model is for the estimation of the marginal process and the relevant covariance function. For the covariance model this represents the time taken to estimate the covariance and to integrate it. The eigendecomposition of any covariance function was not included in this time as it would be performing the same function on the same dimensional object. Overall the marginal process was significantly faster than the marginal covariance averaging values just above 0.2 seconds whereas the covariance model would require 5 seconds or more.

Time To Estimate

Sample Size	***n=50***			***n=100***			***n=1000***		
Noise	0%	10%	20%	0%	10%	20%	0%	10%	20%
D1									
Process	0.066	0.072	0.066	0.077	0.068	0.070	0.248	0.225	0.252
Covariance	2.522	2.858	2.522	2.696	2.445	2.466	5.460	4.990	5.281
D2									
Process	0.077	0.064	0.072	0.067	0.075	0.074	0.242	0.236	0.242
Covariance	2.940	2.507	2.867	2.348	2.580	2.619	5.467	5.305	5.304

Table 5.7: Average computation time for the Marginal Process and Marginal Covariance models over all designs in seconds (1 unit = 1 second).

Overall, the sample size does not have a large effect on the reconstruction ability of the models, with exception of the fact that the FPCA model was able to estimate more components to account for noise in both high noise settings with $n = 1000$.

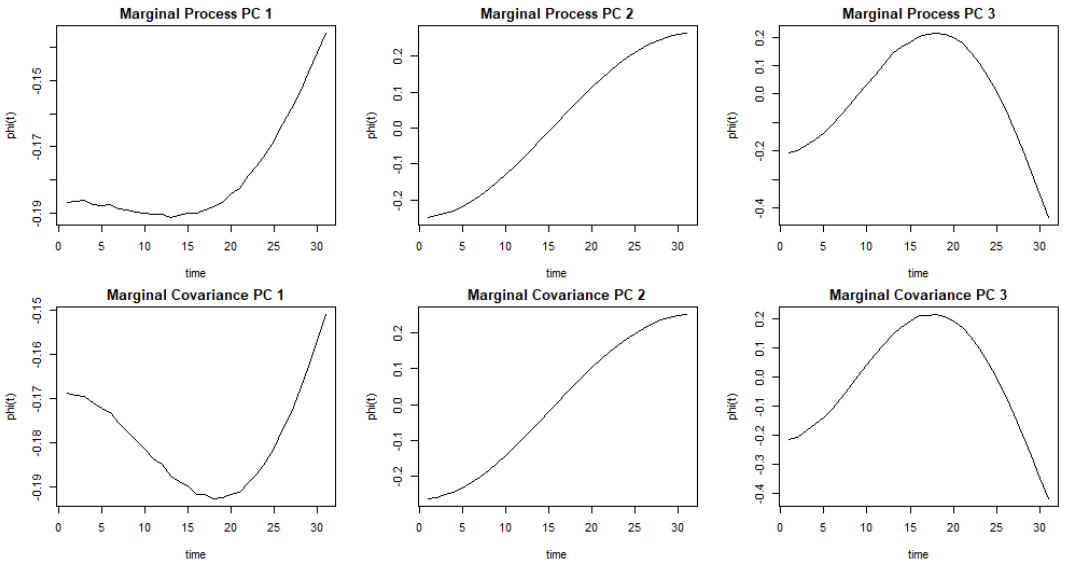


Figure 5.2: D2.2 result: The first three principal components for the decompositions of copies of $X(s,t)$ from D 1.2, simulated with $\varepsilon_{ijk} \sim N(0, 0.2)$. The first row shows $\{\phi_l(s)\}$ and the second shows $\{\varphi_l(s)\}$.

To summarise this section, we could see that overall the highest reconstruction ability was seen in the FPCA model with often the lowest number of components. The marginal covariance and marginal process models had similar performance in Designs 1.1, 1.2 and 2.2. However, the marginal process model was not able to fully recover the variation of the data in Designs 2.1.

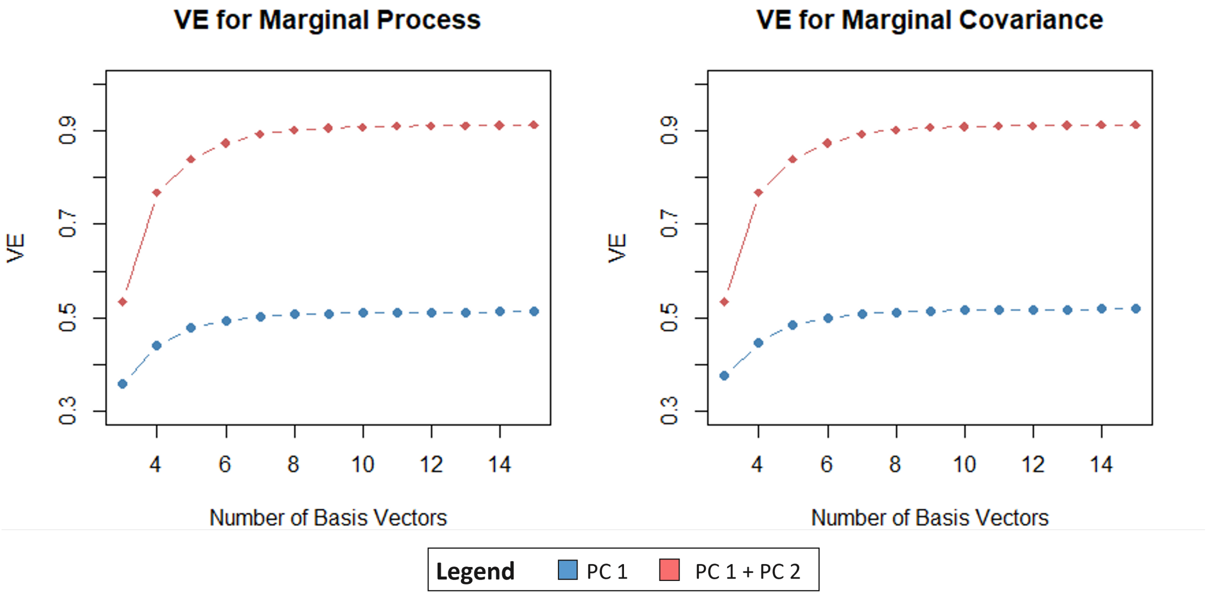


Figure 5.3: Plot of average VE over 100 replicates for different number of b-spline vectors from 3 to 15

Choice of Basis Vector

In this simulation, we set $n = 100$ and $\varepsilon_i(s,t) \sim N(0, 0.16)$ and run the estimation procedure as normal with the exception that the estimated score functions can be formed of $b \in \{3, \dots, 15\}$ b-spline vectors. We will look at how the value of b affects the VE and what would be the optimal choice of b meaning that any additional vector would not increase VE significantly. To define what the optimal number of basis functions was, we denote VE_b as the VE obtained from the PCs and score functions estimated with b b-spline vectors. VE was calculated for each and if $\text{VE}_b - \text{VE}_{b+1} < 0.001$ then b was deemed the optimal number for said bases. For each replicate the vector VE_b and b_{optim} were saved. The results for VE_b are shown in Figure (5.3) for both models. They both have logarithmic forms and for both models the optimal number of b-spline vectors was 9 across all replicates. In both models, there was no significant differences in how the b-splines interfered with the VE.

Influence of Sampling Points

Motivated by the fact that in image processing, it is possible to downsize or downsample images to fit certain memory constraints, we would like to investigate the effect of sampling over s and how it influences PC estimation and the reconstruction of the image. During the simulation, we have kept track of both the discrete and the smoothed versions of the PCs and reconstructions.

Table (5.8) shows the VE for all the simulation settings. In this table, the effect of sampling is less visible than the effect noise has on the reconstruction. Both models appear to have a similar cumulative VE however their VE for PC 1 alone is different, where the difference ranges from 0.005 to 0.021. The highest difference in VE from PC 1 can be seen in the lower right quadrant with lowest sampling and considering the sampled PC. It appears, when comparing the smoothed and sampled columns that smoothing can slightly increase the VE. This improvement is not noticeable for $s_j \in$ with $j \in \{1, \dots, 30\}$ but can be noticed for $j \in \{1, \dots, 20\}$ and $\{1, \dots, 10\}$.

Table (5.9) contains the PC error between the true and the estimated functions for both models. Overall, we can see that as sampling becomes more sparse, the quality of the

	VE Smoothed PC				*VE Sampled PC*			
sampling	**Cov**		**Proc**		**Cov**		**Proc**	
30	**PC 1**	**PC 2**	**PC 1**	**PC 2**	**PC 1**	**PC 2**	**PC 1**	**PC 2**
0%	0.561	0.995	0.555	0.995	0.561	0.988	0.555	0.988
10%	0.516	0.903	0.510	0.903	0.516	0.909	0.511	0.909
20%	0.449	0.783	0.443	0.782	0.448	0.776	0.442	0.776
20								
0%	0.561	0.995	0.555	0.995	0.561	0.971	0.553	0.970
10%	0.516	0.909	0.510	0.909	0.516	0.886	0.508	0.886
20%	0.449	0.784	0.444	0.783	0.448	0.762	0.440	0.762
10								
0%	0.561	0.995	0.555	0.995	0.557	0.862	0.535	0.862
10%	0.517	0.911	0.511	0.911	0.512	0.788	0.491	0.787
20%	0.451	0.787	0.445	0.786	0.444	0.677	0.425	0.676

Table 5.8: Variance explained for different sampling density over s displayed in the first column. 'Smoothed PC' corresponds to PCs fitted with b-splines and 'Sampled PC' refers to the PCs computed from the grid s_j.

estimation goes down. Between sampling density 30 and 20 we see a slight increase in error, whereas the jump is much greater when going from sampling 20 to 10 points. of It can be also seen that the marginal process model has consistently higher errors than the covariance model, this is visible across all settings.

Data Size and Computation

We have previously shown that our proposed model based on the marginal process has a computational advantage compared to more traditional approaches. In this section, we wanted to compare three models with one simulation design, where the two variables that were varied are the sample size and the image size, to better understand the effect of data size on potential model estimation times.

We consider the simplest design, D1 and set the data to have no noise, as noise had little effect on model estimation previously as shown in Table (5.7). We have considered sample sizes ($n \in \{50, 100, 200, 500\}$) and image dimensions ($20 \times 20$, 30×30, 40×40, 50×50). For simplicity, we considered square images with equal number of pixels in the rows and columns: 20, 30, 40 and 50. The increase in total number of pixels is non-linear as the total image size increases exponentially from 20^2 to 50^2.

Table (5.10) shows the results of the simulation with the mean and standard deviation over 100 replicates. The marginal process model was consistently faster than the other

	PC Error Smooth				***Sampled PC Error***			
sampling	Cov		Proc		Cov		Proc	
30	**PC 1**	**PC 2**	**PC 1**	**PC 2**	**PC 1**	**PC 2**	**PC 1**	**PC 2**
0%	0.016	0.052	0.136	0.153	0.016	0.052	0.136	0.153
10%	0.016	0.052	0.137	0.153	0.016	0.053	0.138	0.154
20%	0.016	0.052	0.138	0.154	0.017	0.053	0.139	0.155
20								
0%	0.017	0.093	0.136	0.179	0.017	0.093	0.136	0.179
10%	0.018	0.093	0.136	0.179	0.018	0.093	0.314	0.179
20%	0.018	0.093	0.137	0.180	0.018	0.093	0.138	0.180
10								
0%	0.022	0.213	0.141	0.275	0.022	0.214	0.141	0.275
10%	0.022	0.214	0.141	0.275	0.022	0.214	0.142	0.276
20%	0.022	0.214	0.143	0.276	0.023	0.214	0.145	0.278

Table 5.9: Error between the true and the estimated PCs for different sampling density over s displayed in the first column. ’Smoothed PC’ corresponds to PCs fitted with b-splines and ’Sampled PC’ refers to the PCs computed from the grid s_j.

two models, with the FPCA model being slightly faster and the marginal covariance.

The marginal process model has an average 0.049 estimation time for an image of size 20×20 with $n = 50$. In this case, sample size increases had a small effect on the computation time, with the time rising quite linearly for $n = 100$ (0.094) and $n = 200$ (0.180). A similar trend goes on for other data sizes, but the effect of sample size increases gets larger the larger the image, as expected. In the end, for images of size 50×50, the computation time goes from 0.191 up to 1.256. The increase in image size has a slightly larger effect on computation; in the case of $n = 50$, t goes up to 0.191 but for $n = 200$ it jumps from 0.180 up to 0.561.

The marginal covariance model took significantly longer than the other two models in estimation, with 4.008 at the lowest end ($n = 50, 20 \times 20$) and 239.594 at the highest ($n = 500, 50 \times 50$). In this case, the sample size increases had a relatively models but noticeable increase in computation time, going from 4.008 to 6.358 (20×20) or from 65.237 to 99.180 (40×40). The much more noticeable increase was caused by image size, which follows a pattern that resembles an exponential one. For $n = 100$ the values go from a modest 4.566 up to a 159.792.

The FPCA model falls in between the two previously presented models; however, it remains considerable close to the lower end of estimation times. It again shows steady

image size	**20x20**	**30x30**	**40x40**	**50x50**
n	**Process**			
50	0.049 (0.005)	0.087 (0.005)	0.141 (0.008)	0.191 (0.013)
100	0.094 (0.017)	0.145 (0.016)	0.226 (0.013)	0.287 (0.019)
200	0.180 (0.014)	0.261 (0.021)	0.391 (0.016)	0.561 (0.228)
500	0.195 (0.019)	0.613 (0.027)	0.909 (0.061)	1.256 (0.067)
	Covariance			
50	4.008 (0.083)	20.194 (0.263)	65.237 (0.504)	155.381 (1.818)
100	4.566 (0.112)	21.187 (0.138)	66.529 (0.366)	159.792 (3.453)
200	5.265 (0.179)	23.699 (0.134)	72.584 (1.763)	187.759 (2.835)
500	6.358 (0.082)	30.954 (0.126)	99.180 (3.758)	239.594 (3.821)
	FPCA			
50	0.135 (0.005)	0.281 (0.015)	0.504 (0.039)	0.756 (0.032)
100	0.278 (0.025)	0.546 (0.028)	0.947 (0.041)	1.422 (0.043)
200	0.606 (0.036)	1.155 (0.046)	1.938 (0.063)	3.136 (0.081)
500	1.945 (0.072)	3.084 (0.127)	6.381 (0.247)	9.546 (0.176)

Table 5.10: Computation time (in seconds) taken to estimate three types of models given different sample size n and image sizes. The data was generated following equation for D1. The main values are the average estimation time over 100 replicates and the values in brackets are the standard deviation.

time increases with increases in sample size and larger but reasonable increases with the increase of the images. The sample size increase has a close to linear effect on the time which is most visible in the 30×30 case where the time for $n = 50$ is 0.281 and for $n = 500$ it is 3.084. The change in estimation time with respect to image size appears to be proportional to the number of pixels in the image.

5.3.3 Conclusion

We have introduced existing FPCA models that decompose a random function into the product of two functions spanning different domains. We focused in particular on a model introduced by Park and Staicu, 2015 that computed the PCs from the marginal covariance function. As our aim is to apply FPCA to images, we introduce a model that computes the PCs from the covariance function of a marginal process. This is motivated by computational efficiency and ability to work in high dimensions. We investigate the performance of these models in various conditions and compare them to traditional FPCA.

Overall, we are able to see that when the model fits the data, all three models perform similarly to each other. When the model does not fit the data, the covariance model can compensate and still reconstruct the data well, however, the marginal model might not capture the additional variation. This is likely due to the fact that information can be lost when we integrate at an early stage over the random function which can lead to a loss of information which does not occur with the covariance model. There is potential work in the future to determine this discrepancy and find a limit to the VE lost in such a case. However, the marginal process model can show a sizeable advantage in terms of computational burden as its estimation is less computationally expensive and quicker.

Choosing an appropriate number of vectors to represent the score function has an effect on the reconstruction of the observations so it is still an important step in building the model. A value too high may result in overfitting to the data and does not provide a significant improvement in VE as the relationship between VE and the number of vectors is logarithmic.

We have investigated the sampling density effect on estimation of the PCs and the reconstruction of the image. It appeared that if the sampling had a higher density than the periodicity of the observed wave then it did not impede the estimation of the function. In practice, this would allow us to down-sample an observed function within reason if it allows for computational efficiency.

In conclusion, we have shown that our proposed model works similarly to the Marginal Covariance model when the data fits the assumptions. This provides a computational advantage and allows for the estimation of the model in high dimensional cases. Park and Staicu, 2015 provide a more flexible approach that would work on two dimensional cases, that is more robust against different data structures.

Chapter 6

Extension of STFPCA to Longitudinal Data

6.1 Introduction

Chapter 4 has introduced an FPCA model for high dimensional images over time and Chapter 5 studied its performance and compared it to existing methods in a low dimensional simulation study. Thus far, we have considered functional data captured at regular time intervals. This chapter will focus on a sparse data application to extend this model to longitudinal data, specifically the ADNI dataset introduced in Chapter 3. Images remain fully functional objects, but in contrast to Chapter 4, where the scores were functional, this section will model the scores using a mixed model to find associations between brain regions represented by components and case control status.

From a functional data perspective, models for longitudinal data were originally introduced in the early 2000s. Later publications introduced models for functional data correlated over two domains, with the temporal domain being typically longitudinal. These have been discussed in depth in Chapter 5, and, despite showing promising results, remain limited to lower dimensional data due to computation or methodological challenges (Greven et al., 2010; Gromenko et al., 2012; Gromenko and Kokoszka, 2013; Chen and Müller, 2012; Park and Staicu, 2015). Over the last two decades, an increase in multi-center research initiatives have contributed to the availability of large longitudinal datasets such as ADNI (Mueller et al., 2005) and the UK Biobank (Sudlow et al., 2015; Miller et al., 2016).

We are motivated by the ADNI dataset that follows patients at various stages of cognition at 6 month intervals collecting T1 MRI images and cognitive test results. The onset of Alzheimer's dementia has been associated with accelerated atrophy is several brain regions. Hence, volume reduction in particular brain regions can be considered as an imaging biomarker used to investigate the rate of brain deterioration. A frequent and straightforward approach to model longitudinal progression of patients is proposed in papers such as Guerrero et al., 2016; Mofrad et al., 2021, where brain biomarkers were first extracted and were used to estimate a mixed model. Mixed models continue to be a common approach for longitudinal disease modelling as in addition to MRI biomarkers they can model other covariates. Such methods often involve multiple steps in their pipelines and are reliant on software such as FSL for the determining of regions of interest (ROIs) for further analysis.

In this chapter, we adapt our proposed model from Chapter 4 to longitudinal data, observed on a sparse temporal. First, we investigate the effect of missingness via simulation, then we apply the model to T1 MRIs from the ADNI dataset to find associations between regions of the brain and case control status across patients. The scores obtained at different time points for each subject are assumed to follow a random slope model. The random intercept and the random slope can be interpreted as the subject's status at entry and their temporal effect, respectively. These can be used as summaries of a subject's trajectory over time and can be used to associate it with outcomes via logistic regression. We investigate model fit by subset analysis and end with a discussion as well as suggestions for future work.

6.2 Methods

This section describes the methods used, where the model from Chapter 4 is adapted for a longitudinal case and the scores are no longer random functions. Instead, they form a random vector of observations at different time points that correspond to how each PC is represented in a subject's image at some time point.

Define a random function $X(s,t) \in L^2$ where $s \in S = [0, S_1] \times [0, S_2] \times [0, S_3]$ forms a bounded 3-dimensional space, $t \in \mathcal{T} = [0, T]$ and $S \cup \mathcal{T} \subseteq \mathbb{R}^4$. Let $Y_i(s_j, t_k)$ denote the random variable for subject $i \in \{1, \ldots, n\}$ at time point t_k for $k \in \{1, \ldots T_i\}$, where T_i is the total observations for subject i and for all k, $t_k \in \mathcal{T}$. We assume that at least some subjects i have at least 3 visits, that is $\mathcal{T}_i \geq 3$. Let s_j denote a voxel at point $j = (j_1, j_2, j_3)$ where for all $j, s_j \in S$.

$Y_i(s_j, t_k)$ can be expressed as a function $X(s,t)$ and noise ε_{ijk} in the sample model:

$$Y_{ijk} = X_i(s_j, t_k) + \varepsilon_{ijk}, \tag{6.1}$$

The noise of the i^{th} subject, denoted as ε_{ijk}, is i.i.d. over i, with mean zero and variance σ^2_{ijk} at voxel j and time k. We assume the variance function of $X_i(s_j, t_k)$ is smooth and hence ε_{ijk} has a smooth variance in the neighbourhood of j.

Given the model proposed for $X(s,t)$ and equation (6.1), the random variable Y_{ijk} can be expressed as:

$$Y_{ijk} = \mu(s_j, t_{ik}) + \sum_{l=1}^{\infty} \psi_{ilk}\phi_l(s_j) + \varepsilon_{ijk}, \tag{6.2}$$

where $\mu(s_j, t_k)$ is the mean function, $\phi_l(s_j)$ is an orthogonal basis functions, $\psi_{ilk} = \langle Y_{ijk}, \phi_l(s_j)\rangle$ are the corresponding scores at different time points t_{ik}. Assuming that there exists a number of PCs, L, containing a sufficient amount of variance explained the above expression can be truncated:

$$Y_{ijk} = \mu(s_j, t_{ik}) + \sum_{l=1}^{L} \psi_{ilk}\phi_l(s_j) + \varepsilon_{ijk}. \tag{6.3}$$

In contrast to Chapter 4 where scores were functions $\psi_{il}(t_k)$, the scores ψ_{ilk} form a random vector of length $\mathcal{T}_i$ and follow a random slope model for each $l \in \{1, \ldots, L\}$:

$$\psi_{ilk} = \beta_{0l} + \beta_{1l}t_{ik} + a_{0li} + a_{1li}t_{ik} + \varepsilon_{ij}, \tag{6.4}$$

where β_{0l}, β_{1l} are the population intercept and slope, a_{0li} is a random intercept with variance σ^2_{al0}, a_{1li} is a random slope with variance σ^2_{al1} and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_e)$ is error. The covariance between the random slope and the random intercept is assumed to be zero

$Cov(a_{0li}, a_{1li}) = \sigma_{al01} = 0$, (the covariance of the estimates may be non-zero as we do not enforce ortholinearity; however, it should be negligibly small).

The estimation method for model (6.3) follows the steps outlined in Section 4.3.1 except for the score estimation. Whilst previously the scores were considered functions $\psi(t)$, and hence were estimated using function-on-function regression, here they are scalars observed at separate time points and therefore are estimated using function-on-scalar regression. The estimated scores $\hat{\psi}_{ilk}$ follow a random slope model which is estimated using the `lmer` function from the `lme4` package (Bates et al., 2015) that obtains the estimates using restricted maximum likelihood.

The scores model binary subject outcomes, denoted $C_i \in \{0, 1\}$ via logistic regression. The scores, expressed by equation (6.4), can be interpreted as each subject's trajectory relating to brain regions contained within each PC. Since they can be summarised by the random effects a_{0li} and a_{1li}, these will be used in the logistic regression:

$$\Pr(C_i = 1) = \text{logit}\Big(\alpha_0 + \sum_{l=1}^{L} \big(\alpha_{2l-1} \cdot a_{0li} + \alpha_{2l} \cdot a_{1li}\big) + \alpha_{2L+1}\text{age}_i\Big). \tag{6.5}$$

For large L relative to n, we will use LASSO to identify relevant parameters, which was first introduced in statistics by Tibshirani, 1996. The penalization will only be applied to parameters that correspond to PCs and not age. Let $R_i = -\log\left(\frac{P(C_i=1)}{1-P(C_i=1)}\right)$, then the estimated parameters optimize the following equation:

$$\hat{\alpha} = \min_{\alpha_0, \alpha} \left\{ \sum_{i=1}^{n} \Big(R_i - \alpha_0 - \sum_{l=1}^{L} \big(\alpha_{2l-1)} \cdot a_{0li} + \alpha_{2l} \cdot a_{1li}\big) \Big) \right\}^2 \tag{6.6}$$

$$\text{subject to } \sum_{l=1}^{2L+1} |\alpha_l| < h \tag{6.7}$$

where h is a prespecified free parameter that determines the degree of regularization.

To obtain the penalised estimates we used the package `glmnet` with corresponding publication Friedman et al., 2010. Specifically, the function `cv.glmnet` set to be exclusively LASSO. Implementation was done in `R`.

6.3 Simulation

The purpose of the simulation is to evaluate the performance of the model and estimation method from Chapter 4 in a sparse data setting with various levels of missingness to reflect patients missing appointments over the course of a longitudinal study. We will consider smooth objects in 3 dimensions captured at different time grids with increasing amount of missingness.

6.3.1 Design

Elsewhere in the thesis we have evaluated the effect of the underlying data structure, noise and number of observations on the model, in this case we only focus on sparse time with missingness. We will measure the overall reconstruction ability as well as the ability to estimate the underlying scores and principal components. We assume $\mu(s,t) = 0$ and the principal component number is set to $L = 2$.

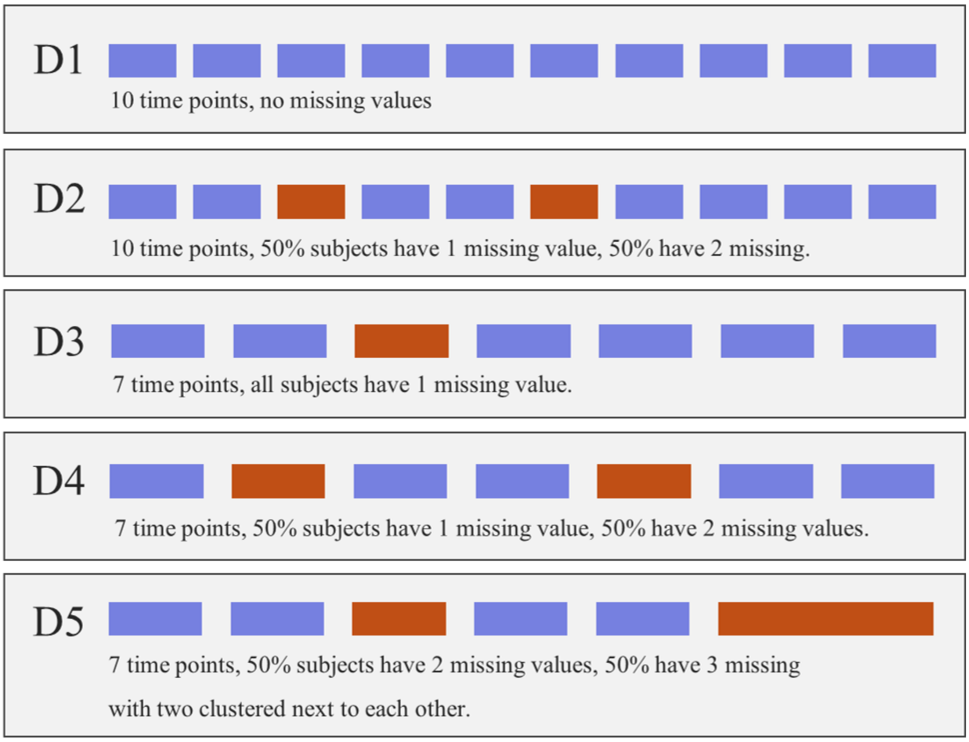

D1
10 time points, no missing values
D2
10 time points, 50% subjects have 1 missing value, 50% have 2 missing.
D3
7 time points, all subjects have 1 missing value.
D4
7 time points, 50% subjects have 1 missing value, 50% have 2 missing values.
D5
7 time points, 50% subjects have 2 missing values, 50% have 3 missing
with two clustered next to each other.

Figure 6.1: An example visualization of the simulation designs, with missing time points being selected at random.

Define the following functions:

$$\psi_{i1}(t) = a_i \cdot \cos(0.5\pi t), \qquad \psi_{i2}(t) = b_i \cdot \sin(\pi t)$$
$$\phi_1(s) = \phi_1(s_1, s_2, s_3) = \sqrt(2) \cdot \cos(2\pi s_1), \qquad \phi_2(s) = \phi_2(s_1, s_2, s_3) = \sqrt(2) \cdot \sin(2\pi s_1),$$

where $t \in [0,1]$, $s \in [0,1] \times [0,1] \times [0,1]$, $a_i \sim N(0,2)$ and $b_i \sim N(0,0.5)$. The functions ϕ_l are evaluated on a $30 \times 30 \times 30$ grid, where $s_{j1}, s_{j2}, s_{j3} \in \{\frac{1}{30}, \frac{2}{30}, \dots, 1\}$, a point on this grid will be denoted s_j. The score functions are evaluated on a grid of K (with $K = 10$ or $K = 7$) equidistant points denoted $t_k \in [0,1]$, specifically $t_k \in \{\frac{1}{K}, \dots, 1\}$. For each simulated series of 3D functions, the number of missing times points as well as the indices k which are missing are chosen randomly. There are five ways this is configured, a quick overview can be found in Figure (6.1). The designs are labelled through $D1$ to $D5$ and each subsequent design involves fewer observations overall. In each case the number of subjects with missing values are predetermined but the subjects are drawn at random. For each subject, then one or two time points are drawn up randomly, in the case of $D5$ the same hold true except one sampling will correspond to two time points being removed at once. Notably, $D5$ is a design that most closely resembles the case found in the ADNI dataset.

The functions form simulated images on a $30 \times 30 \times 30$ grid over K time points using the model equation (6.3):

$$Y_i(s_j, t_{ik}) = \sum_{l=1}^{2} \psi_{il}(t_{ik}) \phi_l(s_j) + \varepsilon_{ijk}, \tag{6.8}$$

where $i \in \{1, \dots, n\}$ for $n = 100$ and $\varepsilon_{ijk} \sim N(0, 0.1)$.

Reconstruction accuracy will be evaluated using variance explained (VE) defined in Chapter 4 as equation (4.12). The estimation accuracy of the components is evaluated using integrated square error (ISE) for the PCS and mean integrated squared error (MISE) for the scores. They are defined as follows:

$$\text{ISE}(\hat{\phi}_l(s)) = \int (\phi_l(s) - \hat{\phi}_l(s))^2 \, ds$$
$$\text{MISE}(\hat{\psi}_{il}(t)) = \frac{1}{n} \sum_{i=1}^{n} \Big[\int (\psi_{il}(t) - \hat{\psi}_{il}(t))^2 \, dt \Big].$$

This will be done per replicate. The results of the simulation will be the mean and the standard deviation of these values computed over the 100 replicates.

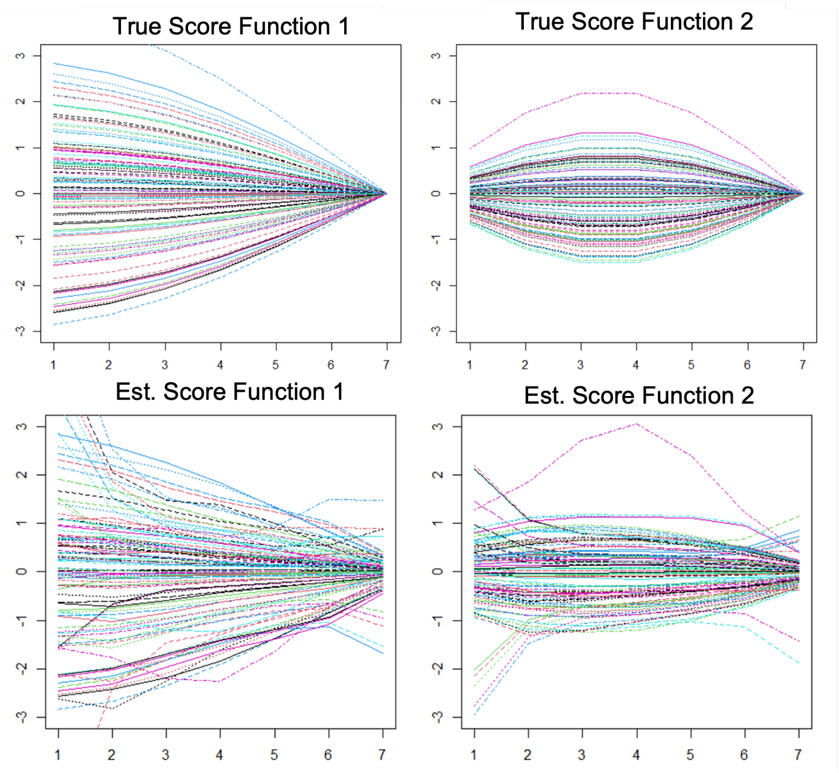


Figure 6.2: D5 simulation reconstruction.

6.3.2 Results

The variance explained and error between true and estimate principal component and scores can be seen in Table (6.1). With the noise included, the total true VE from the two PCs alone was 0.96.

It appears that as the missingness gradually increases, the VE decreases. We can see that as the proportion of observed time points goes down, the total the VE goes from 0.923 to 0.859 ($D1$ to $D5$). Additionally, the standard deviation increases showing that the VE can is more susceptible to variation due to the generated sample.

The principal components appear to be estimated with similar accuracy and are less susceptible to error resulting from missingness. Whilst the errors vary between the different designs there does not appear to be a clear pattern in the average accuracy over the replicates. The design of the principal components is the same as in Chapter 4 and hence a cross-section of the PCs is the same as in Figure (4.1). The scores on the other hand appear to increase in errors in reconstruction as the missingness increases and their standard deviation also increases. The score reconstructions from one replicate of $D5$ can be seen in Figure (6.2). As in the simulation from Chapter 4, the endpoints of the functions are not as accurately estimated; this is likely due to the choice of b-spline basis as those can struggle with the estimation of functions at their end-points. In addition, the subjects whose endpoints happen to be missing have poor reconstructions and can be seen in the figure as deviations from the expected pathway.

	VE		Error			
	$\phi_1(\cdot)$	$\phi_2(\cdot)$	$\hat{\phi}_1(\cdot)$	$\hat{\phi}_2(\cdot)$	$\hat{\psi}_1(\cdot)$	$\hat{\psi}_2(\cdot)$
D1	0.615 (0.03)	0.923 (0.01)	0.611 (0.59)	0.579 (0.54)	0.533 (0.03)	0.377 (0.03)
D2	0.615 (0.03)	0.919 (0.01)	0.627 (0.59)	0.611 (0.56)	0.560 (0.04)	0.392 (0.03)
D3	0.600 (0.03)	0.914 (0.01)	0.617 (0.59)	0.726 (0.55)	0.571 (0.04)	0.424 (0.05)
D4	0.603 (0.04)	0.911 (0.01)	0.697 (0.58)	0.637 (0.57)	0.712 (0.06)	0.526 (0.06)
D5	0.601 (0.06)	0.859 (0.16)	0.702 (0.61)	0.654 (0.57)	0.723 (0.06)	0.539 (0.07)

Table 6.1: Cumulative VE and Integrated Square Error for the estimated functions. The values represent the mean and the standard deviation over 100 replicates.

In conclusion, the simulation has shown that a gradual increase of missing time points reduces our ability to estimate the underlying functions. In particular, subjects with multiple missing time points in sequence have a lowest accuracy in reconstruction, which can be attributed to the smoothing technique not making strong assumptions about the underlying structure of the functions it estimates. A potential avenue to overcome this would be to impute the missing data separately, prior to smoothing the functions. The time invariant functions appeared to be less influenced by the missingness which can be interpreted as the fact that the time-invariant basis can be more easily estimated whilst being less susceptible to error as a result of sparse time points.

6.4 Data Analysis

Alzheimer's Disease Neuroimaging Initiative is a large longitudinal study following patients over the course of years and collecting multiple biomarkers related to the onset and progression of dementia. The data was sourced from the publicly available predetermined ADNI T1 MRI collection. The complete dataset contains 382 subjects across three stages of cognition: CN, MCI and AD. The images were downloaded from the ADNI (loni) website and preprocessed using the pipeline described in Section 3.4.3. The images uploaded to R using the `oro.nifty` package (Whitcher et al., 2011) and were subsequently downsampled to sizes 85 × 105 × 95 and smoothed using a Gaussian blur from the `smooth3D` function from the package `aws` (Polzehl et al., 2020) with the parameter h representing bandwidth set to 0.7 (in the case of Gaussian kernels it is measured in half-power bandwidth).

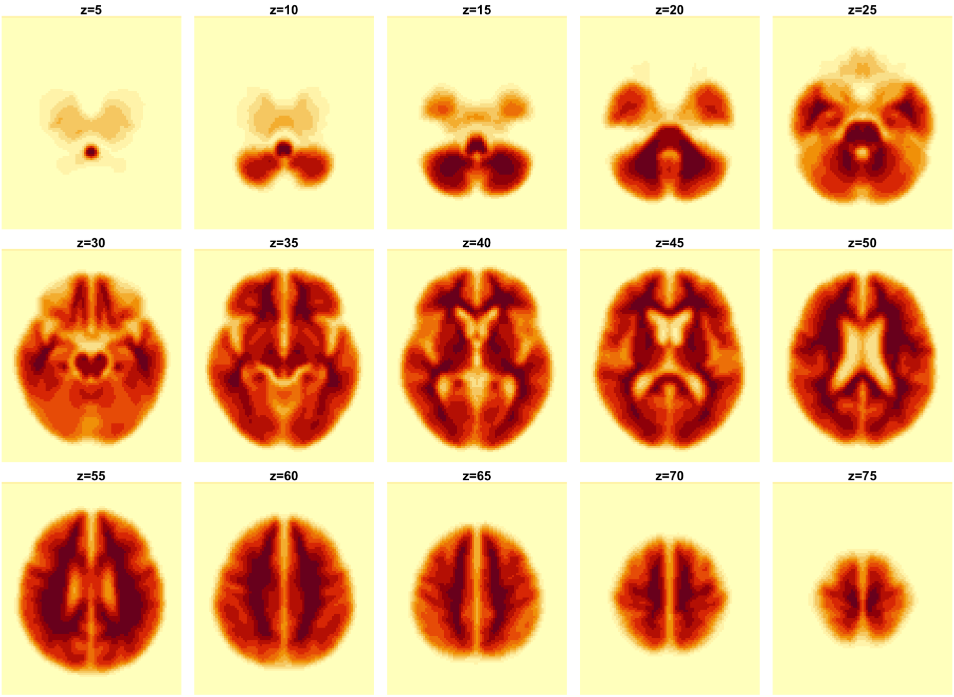


Figure 6.3: Slice of estimated global mean $\mu(s,t)$ for a age restricted subset of 138 subjects.

We fit the model defined in equation (6.3), using the estimation method described in Chapter 4. Like in the case of the simulation, the method was adapted to account for variation in the density function $g(t)$ across subjects. First, the global mean $\hat{\mu}(s,t)$ was computed pixelwise across subjects, the mean at time point 1 is plotted in Figure (6.3). The resulting centered function $V_i(s_j, t_{ik})$ is then marginalised with $V_i(s) = \int_{\mathcal{T}_K} V_i(s_j, t_{ik}) g(t_k) dt$.

Subject image variation $\hat{\sigma}^2_{iqjk}$ is estimated by splitting the image into small $2 \times 2 \times 2$ cubes as described in Section 4.3.1. Principal components are estimated using regression and are subsequently smoothed using 3D penalized smoothing, we adopt the cubic b-spline basis with 42, 52, 45 knots along the x, y, z axes, respectively. The penalty matrices for all axes are of order 2. Tuning parameters in the penalized smoothing are selected by minimizing the GCV values. The scores are updated using function-on-scalar regression per time point.

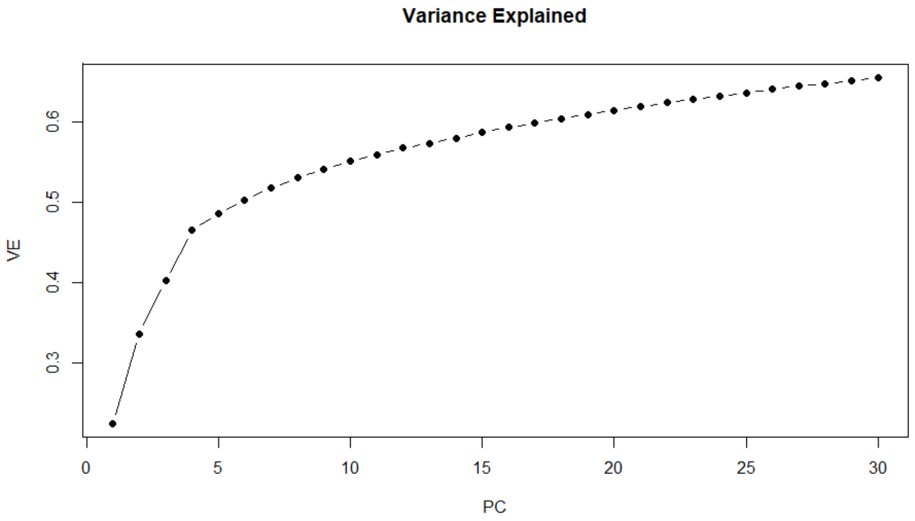


Figure 6.4: Variance explained for the case-control group, the x-axis represents the cumulative number of components.

6.4.1 Case-Control Group

We first consider the cohort of patients who have been consistently diagnosed as CN or AD for the entire duration of the study, assuming that in the visit they have missed they would have had the same diagnosis. Amongst the 382 subjects in the complete cohort, 123 are CN and 96 have an AD diagnosis for the whole study. The total of 219 subjects have at most 5 visits and will be included in the following analysis. This data will be used to estimate the STFPCA model and the scores will be used in a logistic regression model to determine which components have an impact on healthy brain ageing contrasted with dementia.

Results

Cumulative variance explained (VE) for each PC (defined in equation 4.12), is plotted Figure (6.4). Whilst the model is able to identify 218 PCs, the first 10 PCs explain 55% of VE, whereas the set of 30 PCs explains 65% of VE. Slices of the PCs are plotted in Figure (6.5) where one can see that the highest values of PCs are centered around the ventricles. We consider the first 10 PCs as any subsequent PC explains less than 1% of VE. The corresponding scores are modelled with equation (6.4).

To link the subject outcome with the score function, we use subject age as well as the random effects a_{0li}, a_{1li} in a logistic regression from equation (6.5) with $L = 10$. The outcomes are stored as a binary vector where 0 indicates a control and 1 indicates a case. Including the intercept and age, there are 24 parameters to be estimated in the logistic regression. We use LASSO to identify relevant covariates to be included in the final model, with the penalization set to omit the age covariate. From this approach, 11 covariates have been identified that correspond to numerous PCs as seen in Table (6.2) together with the penalised estimates. The scores identified in Table (6.2) are plotted in Figure (6.6).

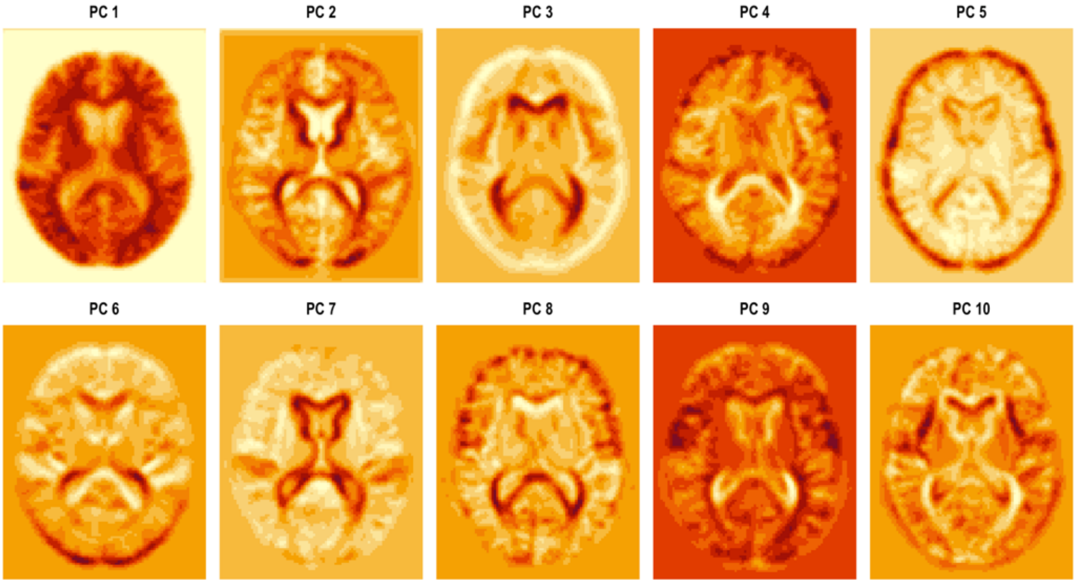


Figure 6.5: Slices along the z-axis ($z = 45$) for the first 10 Pcs of the case-control group (219 subjects).

Looking at Table 6.2, of the selected covariates, 8 correspond to random intercepts and 3 correspond to random slopes. This suggests that overall, the subject state at entry may matter more in terms of determining their outcome. Looking at Figure (6.6), the scores whose slopes were selected are indexed 2, 7 and 9 and it appears that the direction of the slopes for two groups is different in the plotted figure. For example, in score 2, the AD group seems to have a slight slope going upward whereas the CN group is flat. The penalised coefficients for the slopes are larger than the intercepts, which corresponds to the fact that the slopes have small values $a_{1\ell i}$ compared to $a_{0\ell i}$ and age. The influence of age is relatively small, with age being positively correlated to AD. Notably, age is a larger variable in the range of $[60, 95]$ whereas other covariates included in the model are closer in range of $[-1, 1]$. Finally, the coefficients seem to increase as ℓ is higher, which corresponds to the fact that the overall variation of the scores should go down as ℓ increases.

	α_0	age_i	$a_{0\ell i}$	$a_{1\ell i}$	$a_{0\ell i}$	$a_{0\ell i}$	$a_{0\ell i}$
ℓ			1	2	3	4	5
Coef.	-4.46	0.06	-5.82	17.88	11.38	-1.69	1.80
	$a_{0\ell i}$	$a_{0\ell i}$	$a_{1\ell i}$	$a_{0\ell i}$	$a_{0\ell i}$	$a_{1\ell i}$	
ℓ	6	7	7	8	9	9	
Coef.	-14.04	32.14	-162.26	-6.72	-39.56	380.65	

Table 6.2: Penalised coefficients for the logistic regression model in the case-control group

6.4.2 Model Fit

We would like to further investigate model fit as we observe that the amount of variance explained by 20 components is rather low, namely 60%. This suggests that our functional model might not fit well. One reason might be that the images are too different due to variation in severity of disease and in age of the participants. To investigate this, we now consider more homogeneous group by taking a sub-sample of the original 382 subjects, considering only those aged between 69 and 75 at screening. This group contains 138 subjects of which 55 are CN for all time points, 29 are AD and a further 29 are MCI, the remaining subjects change diagnoses from CN to MCI at some points in the study.

Results

Model (6.3) is fitted as in the case-control group. The estimation method was able to identify all 138 PCs with the first 22 explaining 70% of the total variation and the first 50 explaining 80% of the variation. VE (defined in equation (4.12)) for each PC is plotted Figure (6.7), where the x-axis represents the number of PCs and the y-axis is the cumulative VE. The contribution of the first 20 PCs to VE can be seen in Table (6.3). Slices of the PCs are plotted in Figure (6.8).

The scores are modelled with equation (6.4) and the random intercepts and slopes are used to model the outcome variable in logistic regression. We fit the model described in equation (6.5) with age excluded and $L = 22$. As before, we use LASSO to identify relevant covariates from a total of 45. In this approach, 4 covariates were identified as

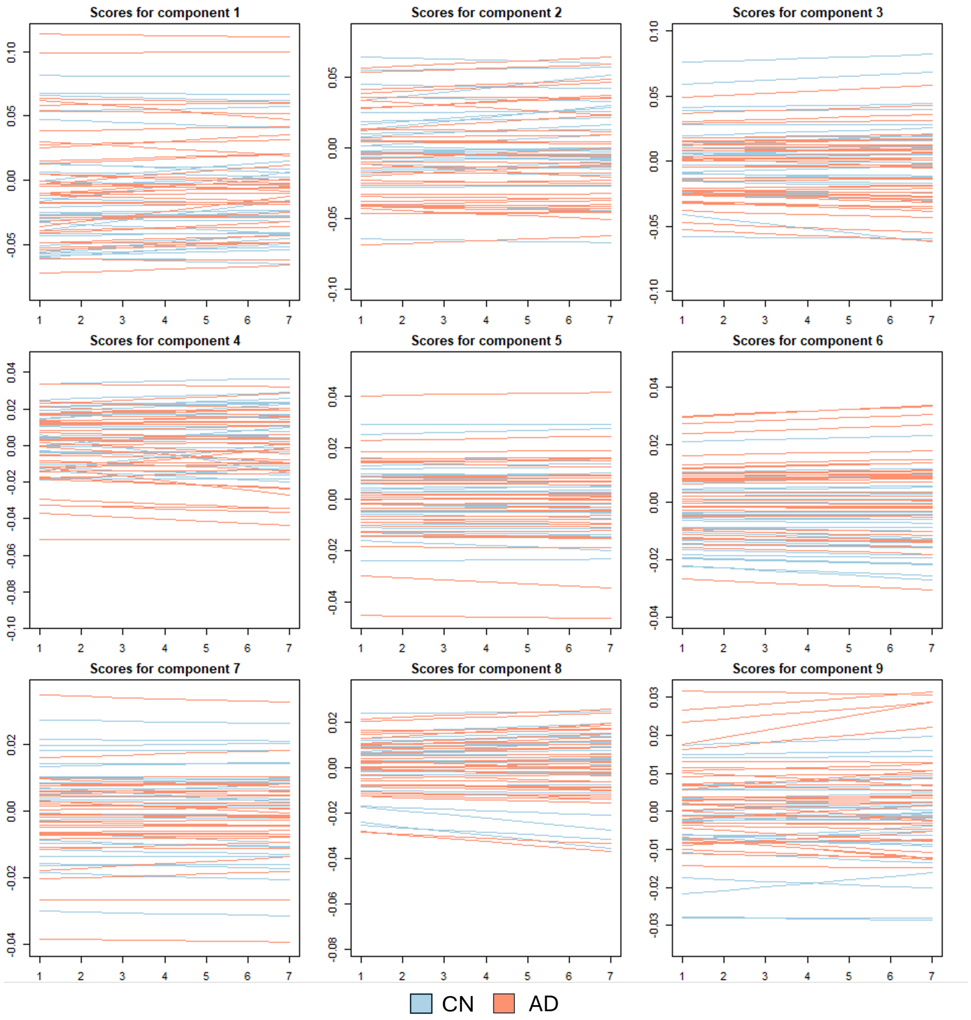


Figure 6.6: Scores for the case-control group estimated from models 6.4 random intercept and random slope a_{0li}, a_{1li} that were identified by the LASSO regression in 6.2.

shown in Table (6.4) which correspond to PCs 1,2,5 and 16. In this case, the random intercepts for PCs 1,5 and 16 were selected and the random slope for PC 2.

In both subject groups, PCs 1 through to 5 appear to be very similar, with their inner product being larger than 0.8. After this the PCs computed appear to be different, with all inner products being below 0.5, which shows how the choice of sample can alter the estimated model. Additionally, logistic regression covariates a_{01i}, a_{12i} and a_{15i} were also chosen for the case-control group so this is a consistent finding across subject groups.

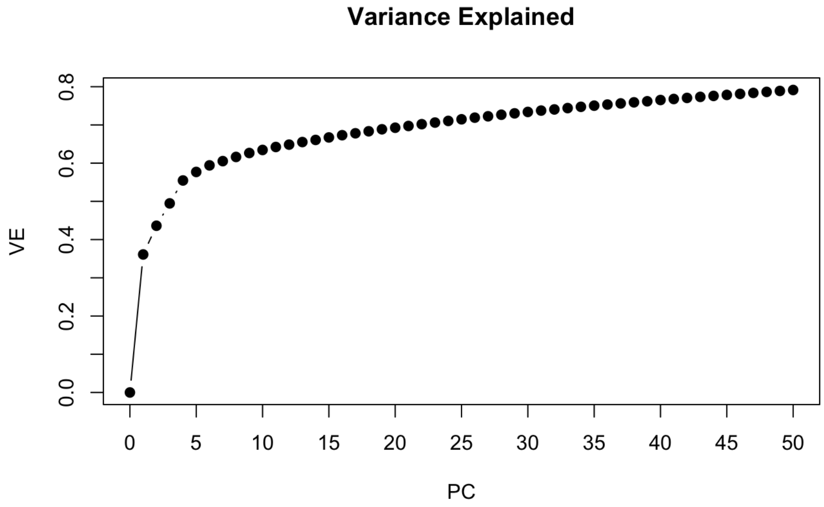


Figure 6.7: Variance explained.

	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_3(s)$	$\hat{\phi}_4(s)$	$\hat{\phi}_5(s)$	$\hat{\phi}_6(s)$	$\hat{\phi}_7(s)$	$\hat{\phi}_8(s)$	$\hat{\phi}_9(s)$	$\hat{\phi}_{10}(s)$
VE	0.361	0.075	0.060	0.058	0.022	0.017	0.011	0.011	0.010	0.008
	$\hat{\phi}_{11}(s)$	$\hat{\phi}_{12}(s)$	$\hat{\phi}_{13}(s)$	$\hat{\phi}_{14}(s)$	$\hat{\phi}_{15}(s)$	$\hat{\phi}_{16}(s)$	$\hat{\phi}_{17}(s)$	$\hat{\phi}_{18}(s)$	$\hat{\phi}_{19}(s)$	$\hat{\phi}_{20}(s)$
VE	0.008	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.004

Table 6.3: The Variance Explained for the first 20 PCs.

6.5 Conclusion and Discussion

Our proposed model and estimation method was able to identify all possible PCs in this scenario, however the amount of variance explained with a reasonable amount of components was dependent on the overall variability between subjects. In the case-control group, the first 10 PCs explained 55% of variation whereas for the smaller, more homogeneous group, the first 10 estimated PCs reached 63% VE. In both cases, PCs after 10 explain less that 1% of the variation. This suggests that the dataset still contains a large amount of intra-subject variability with regards to age, disease state and other factors.

Analysis on the scores identified relevant PCs that are most associated with subject diagnosis. LASSO was used for parameter selection allowing us to identify PCs that are most correlated with subject outcome. The indices of these PCs for both groups are found in Tables (6.2) and (6.4). In both groups, the random intercept of scores 1 and 5, as well as the slope of score 2 were deemed as important predictors of outcome. For scores 1 and 5, this suggests that different state upon entry can differentiate the outcome

	α_0	$a_{0\ell i}$	$a_{1\ell i}$	$a_{0\ell i}$	$a_{0\ell i}$
ℓ		1	2	5	16
Coef.	-0.516	17.419	-5.610	-32.018	-1.44

Table 6.4: Penalised coefficients for the logistic regression model

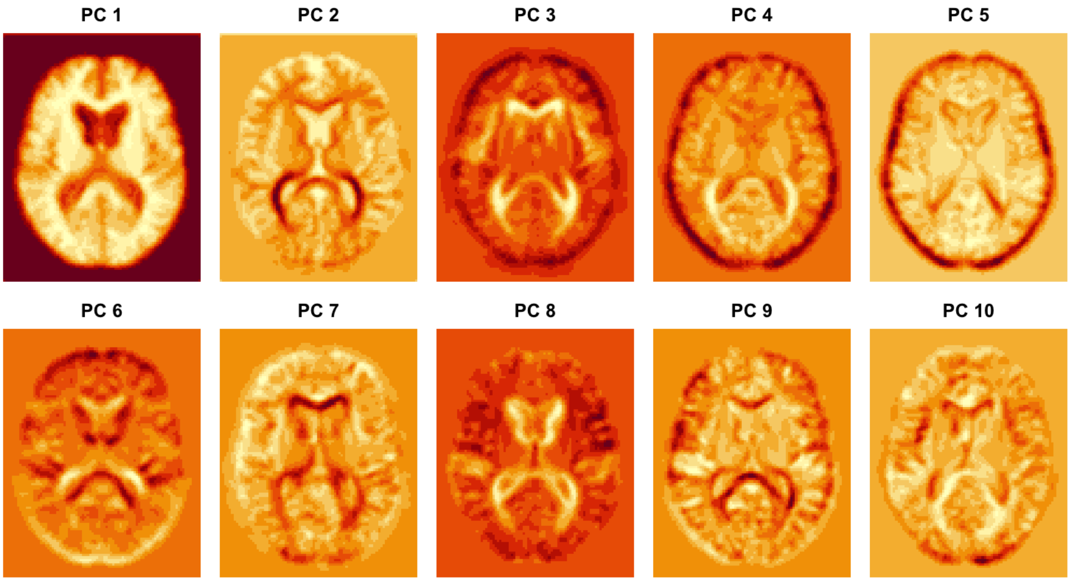


Figure 6.8: ADNI principal component slices along z=45 for the age-subset (138 patients).

whereas for score 2, it suggests that the change over time is indicative of a diagnosis. The PCs indexed 1,2 and 5 appear to correspond to ventricle size and seem to have an effect on the diagnosis in the regression model. This is consistent with previous findings in the literature (Thompson et al., 2004).

Whilst the number of longitudinal observations may be less of a limiting factor in our model, the score functions could be modelled differently in the future. We have chosen a simple approach that assumes the underlying structure of the score functions to be linear but other modelling methods can be considered in the future that may include more flexibility. Milliken and Edland, 2000 point out that the rate of decline is non-linear, so using the diagnosis time or time of enrolment may not directly reflect the patients trajectory. Using a pre-determined function such as a sigmoid or a more flexible model could be more appropriate in the future. We have found that model fit is determined by the heterogeneity of the data, the case-control group was able to reasonably estimate components that explain 55 variation with $L = 10$. A model fitted on a smaller, more homogeneous subject group could recover 63% of VE with 10 PCs, however these are

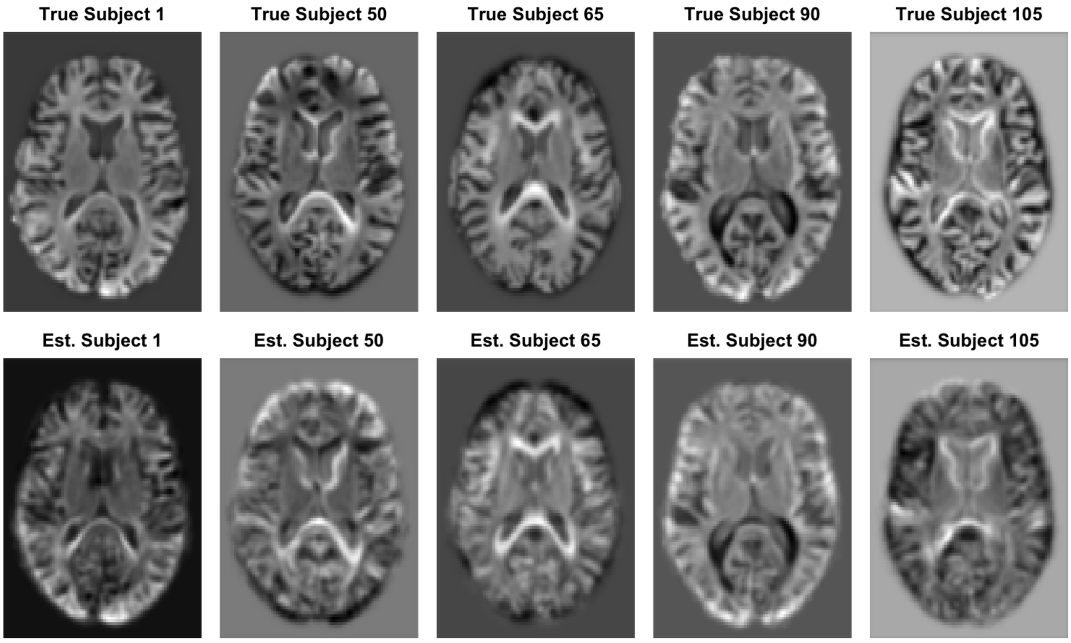


Figure 6.9: ADNI image reconstruction along the axis z=45 and time point 1 for random subjects from the age group.

still relatively low compared to our application in Chapter 4. This could be addressed by potentially considering a large and/or a more homogeneous dataset.

In this chapter we have considered association between brain regions, represented by PCs and binary subject outcomes. Much of the literature in this field, as discussed in Chapter 3, is interested in using imaging for prediction and we would like to investigate our approach in this context.

Chapter 7

Temporal Images for Prediction of Dementia

7.1 Introduction

In all chapters of this thesis, we have considered FDA methods for the analysis of neuroimaging data. Chapter 6 extended the STFPCA model to longitudinal data and was applied to the ADNI dataset, which has greatly contributed to methodological development in multiple fields. Advancements in computation power have fueled the popularity of machine learning techniques to analyse images and to use them to predict outcomes. In this chapter, we are motivated by a method published by Sauty and Durrleman, 2022 that proposed a neural network to learn the representation of longitudinal images. We propose a network inspired by their work and a framework for the comparison of networks to the model discussed in Chapter 6.

Whilst much work has been done to predict disease status from imaging (discussed in Section 3.4.2), relatively few ML methods exist where a series of images over time is considered at once. We are motivated by Sauty and Durrleman, 2022 who propose a variational autoencoder with a mixed model regulating the latent space such that each latent variable represents a feature extracted from the image modelled on a longitudinal trajectory. In contrast to our proposed statistical model, this method could potentially find non-linear relations between voxels of an image and summarise them as a random vector of latent variables. On the other hand, NNs are known for lacking interpretability and can be computationally expensive, with both training and testing taking a lot of computing power.

In this chapter, we suggest an adaptation of the network proposed by Sauty and Durrleman, 2022, that can be more directly compared with our modelling approach from Chapter 6. In our case, we will model the latent variables with a random slope model and use the random intercepts and slope in a logistic regression to predict patient case-control status. As the network is a non-linear approach, we are interested in its performance compared to our model that relies on the linear addition of PCs to reconstruct the image. We suggest a framework using cross validation for the comparison of this network and the STFPCA model that uses the latent variables in logistic regression to predict patient outcomes.

7.2 Background on Variational Autoencoders

Autoencoders are a standard network for non-linear dimensionality reduction, comprised of an encoder $q_\phi(\cdot)$ and a decoder $p_\theta(\cdot)$ that first reduce the input to a pre-determined latent space and use this latent space to reconstruct the original image. Let Y_i, i denoting a subject, be a real-valued vector or matrix that can represent a set of covariates or an image in $\mathbb{R}^d$ with $d \in \{1, 2, 3\}$. The function $q_\phi : Y_i \to z_i$ is parameterised by weights ϕ and maps the input $Y_i \in \mathbb{R}$, to a vector $z \in \mathbb{R}$. The decoder $p_\theta : z_i \to \hat{Y}_i$ maps the latent space back to $\mathbb{R}^3$ to estimate the random variable. The parameters of the network, often called weights, are updated via back-propagation where the loss function is the reconstruction error denoted $\mathcal{L}_{recon}$.

Variational autoencoders (VAE), introduced by Kingma and Welling, 2014, extend the autoencoder by treating the functions q_ϕ and p_θ as estimations of the cumulative distribution function. In this scenario, we assume that the random variables Y are generated from a latent variable z that follows some unknown distribution $p(z)$, and hence they follow a posterior distribution $p(Y|z)$. We assume that both functions can be parameterised by $p_\theta(z)$ and $p_\theta(Y|z)$, and that their PDFs are differentiable almost everywhere w.r.t. both θ and z. However, both θ and z are unknown. This poses an intractable problem, as $p_\theta(Y) = \int p_\theta(z)p_\theta(Y|z)dz$ cannot be evaluated. To circumvent this issue, the authors introduce a recognition model $q_\phi(z|Y)$ which is an approximation of the intractable true posterior $p_\theta(z|Y)$. From a coding perspective, $q_\phi(z|Y)$ is treated as an encoder and $p_\theta(Y|z)$ as the decoder, and the parameters ϕ and θ will be estimated jointly. The loss function includes another parameter $\mathcal{L}_{KL} = D_{KL}\big(q_\phi(z|Y_i)||p_\theta(z)\big)$, which denotes the Kullback–Leibler (KL) divergence.

In traditional VAEs, the latent space Z is made up of continuous random variables z sampled from $q_\phi(z|Y)$. As $q_\phi(z|Y)$ cannot be computed directly, the authors introduce a reparameterization trick. There, the random variable z is treated as deterministic with $z = g_\phi(\epsilon|Y)$ where ϵ is an auxiliary variable with independent marginal $p(\epsilon)$ and $g_\phi(\cdot)$ is a vector valued function parametrised by ϕ. Kingma and Welling, 2014 provide various options of defining ϵ and $g_\phi(\cdot)$. Focusing on their example of a VAE, they define the approximate posterior be a multivariate Gaussian with a diagonal covariance structure:

$$\log q_\phi(z|Y_i) = \log \mathcal{N}(z; \mu_i, \sigma_i^2 I) \tag{7.1}$$

where the mean and s.d. of the approximate posterior, μ_i and σ_i^2, are outputs of the encoding MLP. Then $z_i = g_\phi(Y_i, \epsilon_i) = \mu_i + \sigma_i \cdot \epsilon_i$.

7.2.1 Motivation: Longitudinal VAE

Sauty and Durrleman, 2022 introduce a longitudinal VAE where the latent space elements z follow the mixed effect model:

$$z_{ij} = p_0 + \big[e^{\xi_i}(t_{ij} - \tau_i)\big]v_0 + w_i + \varepsilon_{ij}, \tag{7.2}$$

where the random effects of the model are: e^{ξ_i} and τ_i, denoting the acceleration factor and the onset age of patient i, respectively. Then w_i is the space shift that encodes inter-subject variability. They assume $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tau_i \sim \mathcal{N}(t_0, \sigma_\tau^2), \xi_i \sim \mathcal{N}(0, \sigma_\xi^2)$ and $w_i \sim \mathcal{N}(0, 1)$. The parameters p_0, v_0, t_0 are respectively a reference position, velocity and time and describe the average trajectory. Together with the variances σ_ε^2, σ_τ^2, σ_ξ^2, they form the fixed-effects of the model.

Their model is updated using a composite loss function

$$\mathcal{L} = \gamma_1 \mathcal{L}_{recon} + \gamma_2 \mathcal{L}_{KL} + \gamma_3 \mathcal{L}_{align} \text{ where } \begin{cases} \mathcal{L}_{recon} = \sum_{ij} ||Y_{ij} - \hat{Y}_{ij}||^2 \\ \mathcal{L}_{KL} = \sum_{ij} D_{KL}\big(q_\phi(z|Y_{ij})||N(0,1)\big) \\ \mathcal{L}_{align} = \sum_{ij} ||z_{ij} - \hat{z}_{ij}||^2 \end{cases} \tag{7.3}$$

where $\hat{z}_{ij}$ denotes the estimate outcomes from model (7.2).

The latent variables are parameterised as simple normally distributed variables, i.e. $\hat{z}_{ijk} = \mu_k + \sigma_k \cdot \epsilon_{ijk}$. And hence, the KL loss penalises $q_\phi(z|Y_{ij})$ for not adhering to a standard normal distribution. Simultaneously, the alignment loss penalises the model for producing latent variables that do not adhere to the mixed model. We believe that these two functions cannot be optimised simultaneously as they both assume different things from the latent variables z_{ij}. For this reason, we will introduce a VAE with a simpler model for the latent variables, one that allows us to assume a specific structure on z_{ij} such that both KL and alignment loss have the same optimization.

7.3 Methods

Denote Y_{ij} as a 3D image for subject $i \in \{1, \dots, n\}$ with index j denoting the j^{th} observation for the i^{th} subject with $j \in \{1, \dots T_i\}$. Consider a VAE comprised of an encoder $q_\phi(\cdot)$ and a decoder $p_\theta(\cdot)$. The function $q_\phi : Y_{ij} \to z_{ij}$ is parameterised by weights ϕ and maps the 3-dimensional input Y_{ij} with values $\in \mathbb{R}^3$, to a vector $z_{ij} \in \mathbb{R}^k$, where k denotes the number of latent variables which is chosen prior to training and is set for all subjects. The decoder $p_\theta : z_{ij} \to \hat{Y}_{ij}$ maps the latent space back to $\mathbb{R}^3$ to estimate the image.

Whereas a standard VAE (Kingma and Welling, 2014) assumes all elements of the vector z_{ij} to follow a normal distribution, we propose the following random slope model that models the latent variable as a linear relation between constants, subject identity idexed i and subjects visit time t_{ij}:

$$z_{ij} = \beta_0 + \beta_1 t_{ij} + a_{0i} + a_{1i} t_{ij} + \varepsilon_{ij}, \tag{7.4}$$

where β_0, β_1 are the fixed effects, a_{0i} is a random intercept with variance σ^2_{a0}, a_{1i} is a random slope with variance σ^2_{a1} and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{ei})$ is error. We denote the covariance between the random slope and the random intercept as $Cov(a_{0i}, a_{1i}) = \sigma_{a01}$. We can write the covariance between two observations z_{ij}, z_{hl} as Σ whose number of rows and columns is the total number of observations $n_{tot} = \sum_i^n T_i$ and is defines as:

$$\Sigma = Cov(z_{ij}, z_{hl}) = \begin{cases} 0 & \text{if } i \neq h \\ \sigma^2_{a0} + \sigma_{a01}(t_{ij} + t_{il}) + \sigma^2_{a1} t_{ij} t_{il} & \text{if } i = h, j \neq l \\ \sigma^2_{a0} + 2\sigma_{a01} t_{ij} + \sigma^2_{a1} t^2_{ij} + \sigma^2_{e0} & \text{if } i = h, j = l \end{cases} \tag{7.5}$$

This model still assumes that each $z_{ijk} \in z_{ij}$ follows a normal distribution, however now it has the form:

$$z_{ijk} \sim N(\beta_{0k} + \beta_{1k} t_{ij}, \Sigma_k)$$

where the covariance matrix is structured as above. The subscript k is to denote each element in the latent space as each would have its own model estimated.

The loss function would then have three elements: the reconstruction error $\mathcal{L}_{recon}$, KL divergence $\mathcal{L}_{KL}$ and alignment loss $\mathcal{L}_{align}$, which taken the difference between the true and the estimated latent variables. They are defined as follows:

$$\mathcal{L} = \gamma_1 \mathcal{L}_{recon} + \gamma_2 \mathcal{L}_{KL} + \gamma_3 \mathcal{L}_{align} \text{ where } \begin{cases} \mathcal{L}_{recon} = \sum_{ij} ||Y_{ij} - \hat{Y}_{ij}||^2 \\ \mathcal{L}_{KL} = \sum_{ij} D_{KL}\big(q\mathcal{N}_0 \parallel \mathcal{N}_1)\big) \\ \mathcal{L}_{align} = \sum_{ij} ||z_{ij} - \hat{z}_{ij}||^2 \end{cases} \tag{7.6}$$

Each of these elements are further defined in next subsection.

7.3.1 Training

During training, there will be two phases. In the first phase, we will not be estimating the mixed model and the network will train only with reconstruction loss to establish the initial parameters for the encoder $q_\phi(\cdot)$ and decoder $p_\theta(\cdot)$. In the second phase we will include the KL and alignment loss functions as well as sample z_{ij} following the random slope model.

Prior to training we decide on the number of epochs we would like to train and the number of batches. One epoch corresponds to the full dataset passing through the network and 1 batch is a subset of the data with n_b elements. We denote the set of images in a batch as N_b so that any existing image indexed $(i, j) \in N_b$ means that image i at time j belongs to this batch. Batches are randomised at each epoch so the sets N_b change.

Phase 1 of Training

Phase 1 of training is used only to pre-train the encoder and decoder to stabilise without limitations on the latent space imposed by the other loss functions elements. In this algorithm, we only use the reconstruction component of the loss function and sample the latent variables from a standard normal distribution in the latent space. No KL loss is used to standardise the latent space so the network is allowed to update more freely.

Phase 2 of Training

Once the encoder and decoder are pre-trained, we begin to impose the structure on the latent space. In this section we will discuss how we estimate the random slope model and how the model is used in the network. Here we introduce the model estimation, KL loss and alignment loss.

Estimation of the Random Slope Model

The model from equation (7.4) assumes a structured covariance matrix $\Sigma \in \mathbb{M}_{n_{tot} \times n_{tot}}$, which can be written as a linear combination of matrices multiplied by the covariance parameters. Additionally, as each subject i is independent, we can look at each subject covariance matrix Σ_i that is the covariance between all the observations for subject i

$$\Sigma_i(j,k) = \begin{cases} \sigma_{a0}^2 + \sigma_{a01}(t_{ij} + t_{ik}) + \sigma_{a1}^2 t_{ij} t_{ik} & \text{if } j \neq k \\ \sigma_{a0}^2 + 2\sigma_{a01} t_{ij} + \sigma_{a1}^2 t_{ij}^2 + \sigma_{e0}^2 & \text{if } j = k \end{cases}$$

Then each matrix Σ_i can be written as:

$$\Sigma_i = \sigma_{a0}^2 M_{1i} + \sigma_{a01} M_{2i} + \sigma_{a1}^2 M_{3i} + \sigma_e^2 I \tag{7.7}$$

where $\sigma_{a0}^2, \sigma_{a01}^2, \sigma_{a1}^2$ are the unknown covariance and variances of the random effects and will be estimated during the procedure. The matrices M_1, M_2, M_3 are known $T_i \times T_i$ symmetric matrices. $I \in \mathbb{M}_{T_i \times T_i}$ is a diagonal matrix, M_{1i} is a matrix of 1s and the rest defined as follows:

$$M_{2i}(j,k) = \begin{cases} (t_{ij} + t_{ik}) & \text{if } j \neq k \\ 2t_{ij} & \text{if } j = k \end{cases} \qquad M_{3i}(j,k) = \begin{cases} (t_{ij} \cdot t_{ik}) & \text{if } j \neq k \\ t_{ij}^2 & \text{if } j = k \end{cases}$$

To allow for the estimation of the unknown covariance parameters $\sigma_{a0}^2, \sigma_{a01}, \sigma_{a1}^2$ and $\sigma_e^2\big)$, we can vectorise equation (7.7). Let $vec(\Sigma_i)$ denote the vectorised version of Σ_i where all its columns are stacked into one vector. The same applied to vectorised versions of M matrices, $vec(M_1), vec(M_{2i}), vec(M_{3i}), vec(I_i)$. Then equation (7.7) can be written as:

$$\begin{aligned} vec(\Sigma_i) &= \sigma_{a0}^2 vec(M_1) + \sigma_{a01} vec(M_{2i}) + \sigma_{a1}^2 vec(M_{3i}) + \sigma_{e0}^2 vec(I_i) && (7.8) \\ &= \left(\sigma_{a0}^2, \sigma_{a01}, \sigma_{a1}^2, \sigma_{e0}^2\right)^T W, && (7.9) \end{aligned}$$

where $W = \big(vec(M_1), vec(M_{2i}), vec(M_{3i}), vec(I_i)\big)$ is an $n_{tot} \times n_{tot}$ matrix.

We want to estimate $\beta_0, \beta_1, \sigma_{a0}^2, \sigma_{a01}^2, \sigma_{a1}^2, \sigma_e^2$. We will use iterative Generalised Least Squares:

1. Estimate $\hat{\beta}_0^k, \hat{\beta}_1^k$ using OLS that assumes $\Sigma = \sigma_e^2 I_n$.
2. Compute the residuals $\tilde{z}_{ij}^k = z_{ij} - \hat{\beta}_0^k - \hat{\beta}_1^k t_{ij}$.
3. Compute the product matrix $\tilde{z}\tilde{z}^T$, with $\mathbb{E}[\tilde{z}\tilde{z}^T] = \Sigma$ where Σ is a structured covariance matrix composed of sub-matrices Σ_i along the diagonal.
4. Express the relationship as a linear regression between $vec(\tilde{z}\tilde{z}^T)$ and σ_{a0}^2, σ_{a01}, σ_{a1}^2, σ_e^2. So $vec(\tilde{z}\tilde{z}^T) \sim (\sigma_{\alpha 0}^2, \sigma_{a01}^2, \sigma_{a1}^2, \sigma_e^2)^T W + e$ with W being a $n \times 4$ known matrix defined in equation (7.8).

5. Estimate using GLS $(\sigma^2_{a0}, \sigma^2_{a01}, \sigma^2_{a1}, \sigma^2_e)^T = \left(W^T W\right)^{-1} W^T vec(\tilde{z}\tilde{z}^T)$.

6. Use $(\sigma^2_{\alpha0}, \sigma^2_{\alpha01}, \sigma^2_{\alpha1}, \sigma^2_e)$ to create the structured matrix $\hat{\Sigma}$ defined as in equation (7.5).

7. update $\hat{\beta} = \left(X^T \hat{\Sigma}^{-1} X\right)^{-1} X^T \hat{\Sigma}^{-1} z$ where z is the vector of $z_i j$ and X is the matrix of covariates.

Loss Functions

We would like to use KL loss that is specific to the normal distribution assumed by the random slope model. The alignment loss would be defined as before.

Given two multivariate normal distributions of n_b length random vectors, $\mathcal{N}_0, \mathcal{N}_1$ with separate means μ_0, μ_1 and covariance matrices Σ_0, Σ_1. Their KL divergence can be written as:

$$D_{\text{KL}}\left(\mathcal{N}_0 \parallel \mathcal{N}_1\right) = \frac{1}{2}\left(\operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) - n_b + \left(\mu_1 - \mu_0\right)^{\mathsf{T}} \Sigma_1^{-1}\left(\mu_1 - \mu_0\right) + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right) \tag{7.10}$$

In the case of the network, N_0 would be the distribution obtained by computing the covariance matrix from $q_\phi(z|Y_{ij})$ (aka. the z_{ij} straight from the encoder). Then this would be compared against N_1 which would have the covariance matrix from the estimated random slope model. The KL loss would enforce a covariance structure on the random variables without enforcing the mean.

The alignment loss is defined as:

$$\mathcal{L}_{align} = \sum_{ij} ||z_{ij} - \hat{z}_{ij}||^2 = \sum_{ij} ||z_{ij} - \hat{\beta}_0 + \hat{\beta}_1 t_{ij} + a_{0i} + a_{1i} t_{ij}||^2 \tag{7.11}$$

where couples $(a_{0i}, a_{1i})^T$ are sampled from a multivariate normal distribution with mean zero and 2×2 covariance matrix $[(\hat{\sigma}^2_{a0}, \hat{\sigma}_{a01}), (\hat{\sigma}_{a01}, \hat{\sigma}^2_{a1})]$. The alignment loss would impose the complete structure from the model including the mean $\hat{\beta}_0 + \hat{\beta}_1 \cdot t_{ij}$ and the structured covariance $\hat{\Sigma}$.

Putting this all together, the reconstruction loss ensures that the estimated image is similar to the input, the KL loss enforces the covariance structure on the latent variables and the alignment loss imposes the whole random slope model.

7.4 Data Application

Images are downsampled to $45 \times 50 \times 45$ as 3D convolution layers have many parameters and a large size of a 3D image would increase the size of the network substantially.

The network was pre-trained using just the reconstruction loss for 30 000 and then the mixed model was used. The results shown are after 45k epochs in total. the number of latent variables is set to 32.

7.4.1 Image Reconstruction

The trained network was frozen, meaning that after any forward pass the weights would no longer be updated through back-propagation. Images of certain subjects were passed through to visually evaluate their accuracy as during the whole process, the reconstruction loss should ensure that each subject's reconstructed image should resemble the observed one.

The original and reconstructed images for one subject over time can be seen in Figure (7.1). Images for two different subject are presented in Figure (7.2). In each case, it appears that the reconstructions are very similar and do not differentiate in time or between subjects.

7.4.2 Discussion

The reconstructions have not worked yet on this case. There are many further approaches that could mitigate this issue. Firstly, we use a relatively small dataset considering the number of parameters in a 3D convolutional neural network, and hence the network might not be able to fully learn the variations between subjects. Additionally, the issue may lay within the network (encoder and decoder itself), we have seen that the model was able to converge to the latent values correctly and estimate them relatively well (loss went down) but that did not directly translate to the reconstructions being sufficiently different between subjects.

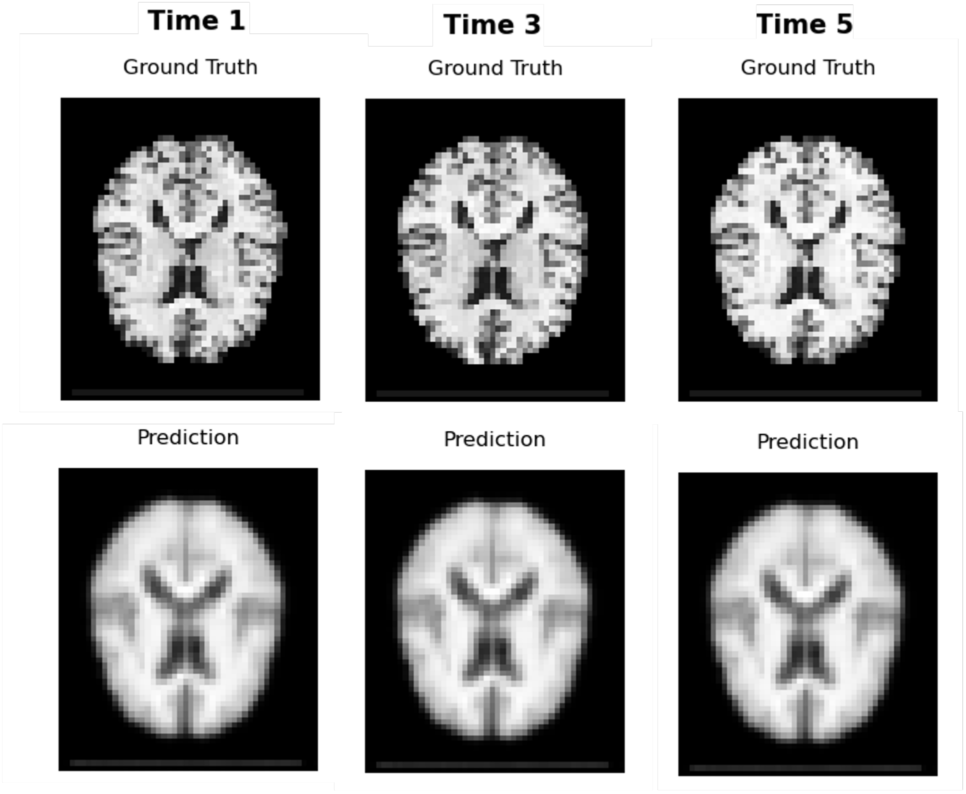


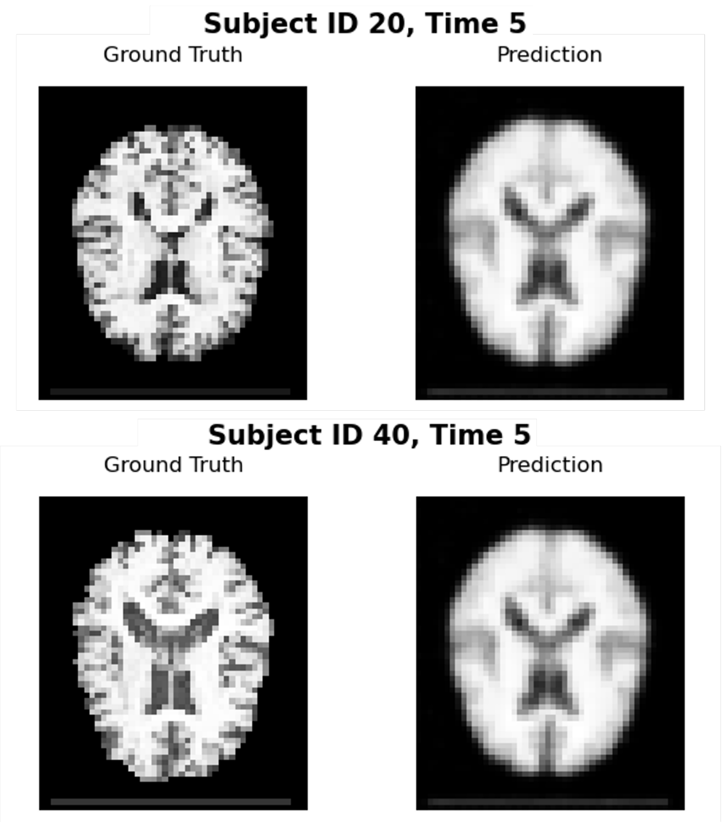
Figure 7.1: Output images from VAE for one subject at three time points, slice along the z axis for $z = 24$.

Another issue may lie within the loss function itself and how each element affects the overall training of the network. A potential step would be to do a large grid search with various settings of hyperparameters $\gamma_1, \gamma_2, \gamma_3$ to see which combination would have a positive effect.

A lot more computational energy goes into training such a network, computing power is much more extensive and fine tuning all the elements together is often done by trial and error, in this case it took a lot of training iteration just to debug the code and to get it working fully we would need to understand the relations between all the hyperparameters as well as find a larger dataset for this specific application. If we do 3D convolutions on 382 subjects it's difficult to evaluate our method working at all.

7.5 Cross-Validation

We propose a framework to compare the FPCA model from Chapter 6 and the aforementioned LVAE network, as both methods propose vary different modelling approaches to a



Subject ID 20, Time 5

Subject ID 40, Time 5

Figure 7.2: Output images from VAE for two subjects at time point 5, slice along the z axis for $z = 24$.

similar dataset. On one hand, the FPCA model is a suitable approach for small dataset with linear assumptions in the decomposition. On the other hand, neural networks, which have increased in popularity in the last decade, provide a flexible approach allowing for the finding of non-linear relations between observed voxels but often can be difficult to train and require large amounts data.

Both the FPCA model and the Longitudinal VAE aim to estimate subjects specific summary variables that can subsequently be used in prediction and reconstruction of the original image. In the case of the FPCA model, it's the score functions modelled with a random slope producing a random intercept and a random slope that summarize the patients trajectory over time. With the VAE, the encoder network is trained to produce latent variables that follow a random effect model as well. As a result, both methods produce a set of random variables which can be compared via the random effect model. Then the parameters of this model can be used in a logistic regression for prediction.

This section aims to evaluate both methodologies with regards to their predictive performance of both models with cross-validation. We consider the case-control group from Chapter 6 that consists of 219 subjects where 123 are diagnosed as CN and 96 as AD. This group naturally lends itself for the task where the models will predict the case-control status of each subject. Due to the potential heterogeneity of the data discussed in 6, we consider a leave-one-out cross-validation (LOOCV) approach, to minimise the effect of large groups missing on the performance.

The score functions from the FPCA model, $\psi_{il}(t_{ik})$, for each $l \in \{1, \ldots, L\}$, follow the random slope model defined as:

$$\psi_{il}(t_{ij}) = \beta_{0l} + \beta_{1l} t_{ij} + a_{0li} + a_{1li} t_{ij} + \varepsilon_{ij}, \tag{7.12}$$

where β_{0l}, β_{1l} are the fixed effects, a_{0li} is a random intercept with variance σ^2_{al0}, a_{1li} is a random slope with variance σ^2_{al1} and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_e)$ is error. The covariance between the random slope and the random intercept is assumed to be zero.

Similarly, each latent variable z_{ijk} for $k \in \{1, \ldots, K\}$ denoting the number of latent components, is modelled as

$$z_{ijk} = \beta_{0k} + \beta_{1k} t_{ij} + a_{0ki} + a_{1ki} t_{ij} + \varepsilon_{ij}, \tag{7.13}$$

where β_{0k}, β_{1k}, a_{0ki} and a_{1ki} are defined as above.

In both cases, the logistic regression that will be estimated in the cross-validation will have the form

$$\Pr(C_i = 1) = \operatorname{logit}\Big(\alpha_0 + \sum_{q=1}^{Q} \big(\alpha_{2q-1} \cdot a_{0qi} + \alpha_{2q} \cdot a_{1qi}\big) + \alpha_{2Q+1} \mathrm{age}_i\Big). \tag{7.14}$$

where q, Q can either be equal to k, K or l, L depending on whether this is for the VAE or FPCA model. And parameters selected using LASSO with age being omitted from penalization.

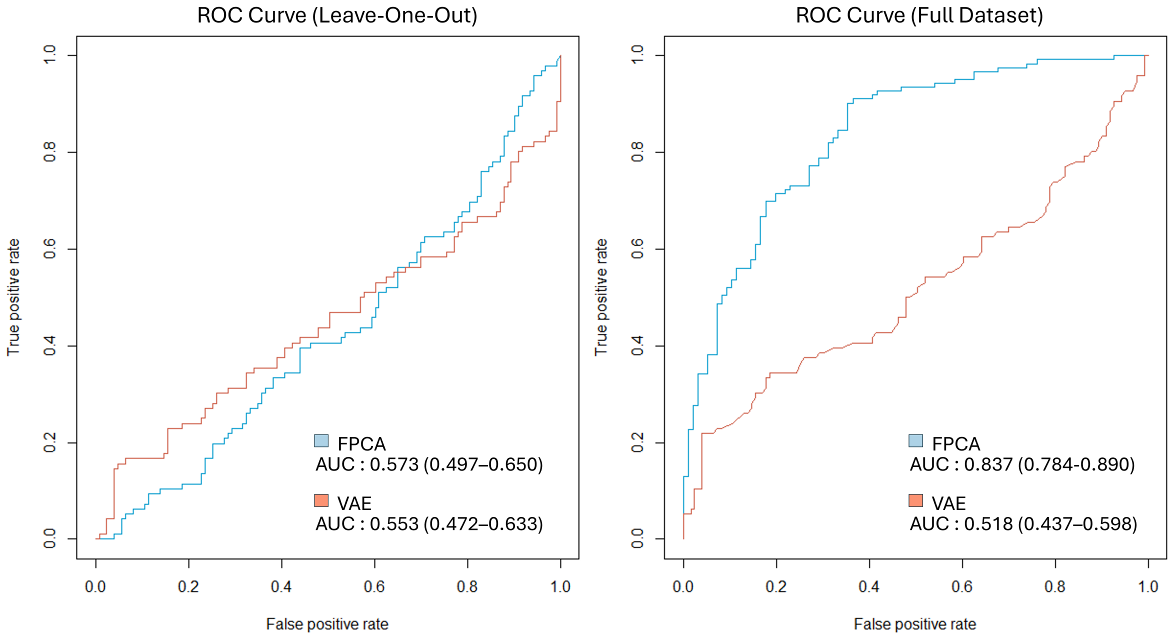


Figure 7.3: ROC Curves for the FPCA model and the VAE, the left are curves from leave-one-out CV and the right are the curves from predictions where the whole dataset was trained.

7.5.1 Design

FPCA

The FPCA model was estimated for the complete 219 subjects and the random slope model was estimated for the score functions. This is considered the complete dataset that will be used in leave-one-out CV. For each set of train subjects, a new LASSO model was estimated to model their outcome as a result of their random intercepts/slopes from eq. (7.12). The test group consists of only one subject for whom the outcome is predicted and will be compared to their true diagnosis. This process is repeated until all subjects have had their outcome predicted.

	α_0	age_i	$a_{0\ell i}$	$a_{1\ell i}$	$a_{0\ell i}$	$a_{0\ell i}$	$a_{0\ell i}$
ℓ			1	2	3	4	5
Avg. Coef.	-4.46	0.06	-7.52	-138.65	13.08	-6.47	4.51
sd.	0.145	0.002	0.269	15.381	0.497	0.683	0.883
	$a_{0\ell i}$	$a_{0\ell i}$	$a_{1\ell i}$	$a_{0\ell i}$	$a_{0\ell i}$	$a_{1\ell i}$	
ℓ	6	7	7	8	9	9	
Avg. Coef.	-22.74	35.63	-924.79	-18.72	-54.70	884.48	
sd.	0.900	0.993	61.486	1.379	1.197	33.120	

Table 7.1: The average estimated coefficient over the course of LOOCV and their standard deviation.

LVAE

The LVAE network was trained on all 382 subjects in the dataset as it needs a larger data size. The latent variables as well as the random slope and random intercept estimated from the IGLS method are extracted for the 219 subjects. These are used in the CV in the same manner as for FPCA, with a new logistic regression model being estimated with LASSO used for parameter selection.

7.5.2 Results

The results of the CV for both methods are shown using ROC curves in Figure (7.3). The first graph represents the prediction accuracy for the LOOCV where the prediction for the test group, consisting of one patient, is compared to their true diagnosis. For this case the AUC is 0.573. The accuracy for the complete model is 0.75 and for the LOOCV the accuracy was 0.43.

The estimated coefficients for the model estimated on the full dataset can be found in Table (6.2), for the LOOCV, Table (7.1) has the mean and the standard deviation of model coefficients over the 219 different train/test splits. Looking at these tables, it can be seen that some coefficients are quite similar across different folds, especially those that correspond to the intercept, age_i and a_{01i}. These coefficients also seemed to be most stable across the folds. Coefficient corresponding to a_{03i} could be considered within reasonable range of the original, but on the whole, the later coefficients have larger estimates than those presented in Table (6.2). Coefficients that correspond to random effects are very high and have the highest standard deviation, which could correspond to model instability across the folds and the drop in AUC in LOOCV. Other coefficients that have a higher ℓ have slightly higher standard deviation, but in general the s.d. seems to be lower than 1 for $\ell < 7$ and for $\ell = 8$ or 9 it is 1.379 and 1.197, respectively. A test LOOCV was ran with the covariates that correspond to the slopes omitted and this did not result in a significant change in AUC.

The second graph in Figure (7.3) represents the ROC curve of the predictions from the logistic model trained on the full 219 subjects with an AUC of 0.837 for the FPCA model and 0.518 for the VAE. For the VAE, the ROC curves were the same for the full estimated model and the LOOCV approach and had a low accuracy of 0.44.

7.6 Discussion

The results for the FPCA model suggest that the data is very heterogeneous, especially when LOOCV was used and one subject missing can largely affect the resulting AUC, going from 0.837 to 0.578, and the accuracy, lowering it from 0.74 to 0.43. For the LVAE, we have seen previously that several elements could be improved in the training that could affect both the reconstruction and prediction outcomes. As it stands, one of the limitations of this project was the data size (for the VAE) and the heterogeneity of the data which affects both methods. This can be seen with how coefficients vary across different folds of the LOOCV with only a few corresponding well to the first model estimated on the full dataset.

Given the current settings of the comparison, the FPCA model performed better but it does not necessarily rule out the network approach. This comparison highlights where each approach could be used and what its advantage and disadvantages are. As stated throughout the thesis the functional approach is an efficient and interpretable approach that could be applied to small datasets. In Chapter 6 we have shown that the heterogeneity of the data affects number of PCs required to fully reconstruct the data, but the method was able to identify PCs which correlate most to case/control status. In contrast, the neural network has the benefit of finding non-linear relations between observations and can be applied to large datasets. This application may have been limited by the data in both approaches, and future work can be focused on incorporating more data into the analysis protocol.

This framework can be improved upon by training the individual FPCA and LVAE models on different train/test datasets, but as this is a demonstration of the method and the network needs further fine-tuning, this approach introduces a feasible comparison in the future for prediction.

A very palpable element that differentiates deep learning from statistical analysis is the requirement for computation. Although neural networks are widely used for the analysis of large medical image datasets, its usage is extremely energy-intensive (García-Martín et al., 2019; Georgiou et al., 2022). Apart from its massive financial cost, it incurs high carbon emissions Strubell et al., 2019. This is visible in the suggested actions to improve model performance in Section 7.4, where many steps are based on observations and iterations. This is contrasted to the statistical approach, where the analysis plan can be planned with minimal experiments prior.

Indeed the network cannot be directly compared to our model as several improvements can be implemented in the future, however this framework can be used for the comparison once the LVAE is trained on a sufficiently large dataset with appropriate hyperparameters chosen. Nevertheless, we have shown that the FPCA model does well on small data and can provide many insights when considering prediction and association, which are harder to infer from network performance during training.

Chapter 8

Conclusion and Future Work

This chapter summarizes the main results of the thesis and includes suggestions, which could improve some aspects of the research. Some future works are proposed as offshoots of the research.

In summary: Chapter 1 gave a general introduction of and a motivation for the work presented in this thesis. Chapter 2 provided background on FDA methods specifically FPCA, functional regression and smoothing. Chapter 3 described the foundations of neuroimaging data, its standard pre-processing steps and introduced the two datasets that are used throughout the thesis. Chapter 4 introduced our proposed spatio-temporal FPCA model and its estimation method which was evaluated in a simulation study. The model was applied to a dense temporal dataset of fMRI images, compared to existing methods and used in a association analysis between active brain regions and risk preference of the subjects. Chapter 5 described a simulation study to compare our model in low dimensions to other methods that would otherwise not be applicable to imaging data. Chapter 6 extended our model to a longitudinal data with sparse time points and missingness. The score functions are modelled with a random slope model which can be used in further analysis. The effect of missingness on model estimation is studies via simulation and the model was applied to the ADNI dataset to find association between brain regions and case/control status. Chapter 7 proposed a novel neural network architecture inspired by previous publications and suggests a framework to compare two methods on the same longitudinal dataset. We show a preliminary use case of this approach to evaluate the performance of the FPCA model and the neural network in terms of case/control status prediction. We discuss further improvements as well as the effect of data heterogeneity on results in Chapters 6 and 7.

8.1 Summary of Results

From this thesis, we make the following conclusions:

1. In Chapter 4, we have shown that our proposed model performs well on densely captured neuroimaging data, as it is able to recover high VE (90.1% with 14 PCs) and the score functions can be used to associate ROIs to subject risk prevalence. The estimation method is robust and allows for the estimation of our model in high dimensions. In a simulation study, we have shown our proposed estimation method can recover underlying functions with high accuracy. Additionally, we have shown our method outperforms existing approaches, specifically one proposed by Li et al., 2019, in both computational efficiency and VE for our dataset.

2. Chapter 5 designed a simulation study in low dimensions to show that our proposed FPCA model fits well to multiple datasets and is comparable to existing state-of-the art models available for lower two dimensional data. We show the computational advantage of our approach and discuss our method's limitations when the underlying model generating data does not fit the estimated model. For cases where the estimated model matched the underlying one, our performance measured by VE and MSE of parameters was the same as other methods. For designs where the underlying model did not match the estimated one, we still recovered a sufficiently large VE with a lower bound of 75% and each of the PCs estimated by our model had the same VE as other methods.

3. Chapter 6 has shown our method can be extended to sparse longitudinal data. A simulation study has shown that our estimation method can recover underlying functions when the observed data has missingness. Application of our model to the ADNI dataset was able to show associations between brain regions whose atrophy has been previously correlated with the onset of dementia. Chapter 7 has shown

that the model has potential to be used in prediction, but the heterogeneity of the dataset used affects the results of CV.

4. Chapter 7 has shown the advantages of using statistical methods on imaging data compared to machine learning methods. Firstly, statistical approaches show an advantage in analysing small datasets and allow for multiple types of analysis which include association, prediction or classification. It discusses the computational efficiency of model estimation compared to network training and hyper-parameter searching, which is both time and energy consuming.

8.2 Publishable Material

The publishable material from the thesis is as follows:

1. Title: ***Spatio-Temporal Functional Principal Component Model for fMRI Data***. This paper is fully written and ready for review, this would be based on the work presented in chapters 4 and 5. This would introduce the new model and estimation method, show its performance via simulation and apply it to our described dataset with association analysis.

2. Title: ***Functional Principal Component Analysis Model for Longitudinal MRI Data***. This paper would be based on chapters 6 and 7. It will contain our proposed model, its extension for longitudinal data, the simulation study measuring the impact of missing data on the estimation and the data analysis with both association and cross validation results.

3. Title: ***Comparison of Statistical and Machine Learning Methods on Image Datasets***. This paper would be extension on the work presented on chapter 7 with a fully trained network and using the framework proposed to evaluate our FPCA model against a VAE.

4. Title: ***Longitudinal Variational Autoencoder for T1 MRI Scans***. This paper would be suitable for a computer science conference presenting the network from chapter 7 trained on a larger dataset with cross validation work to evaluate how well it can predict patient disease status from images.

8.3 Improvements and Future Work

Chapters 5 and 6 have demonstrated the importance of model fit. Firstly, Chapter 5 has shown how the underlying data structure can affect model estimation and the resulting VE. Secondly, Chapter 6 has shown how data heterogeneity can have an effect on the number of PCs required to reach a reasonable VE. Future work can study these phenomena, understand when they occur specifically and potentially find upper bounds for reconstruction ability for cases such as those in design 2 seen in Chapter 5.

In the low dimensional simulation in Chapter 5 we have shown that for the cases where the underlying model does not fit the estimated one, Park and Staicu, 2015 can recover a higher VE with the estimation of additional PCs. This method could potentially allow to model heterogeneous data, however its implementation on imaging datasets in non-trivial. A potential implementation could be based on the representation of images using b-splines and deriving the structure of the covariance matrix as shown for a simpler case in Chapter 8 of Ramsay and Silverman, 2005. This application could potentially address the heterogeneity seen in the ADNI dataset and would provide an extension of the low-dimensional model to images.

In Chapter 6 the score functions were not dependant on age, which was shown to be a factor in the data due to the wide age range upon screening. Further work would be to treat the score function not as beginning from time of screening but rather from the subjects age. This could potentially address some of the heterogeneity issues we have found and would allow for the usage of the full ADNI dataset.

Chapter 7 has presented a novel architecture for a neural network and a framework for comparison, however the network was not trained fully to make a fair comparison. Further work will be done in accordance with the discussion notes in this chapter to produce the aforementioned publications. These are mainly related to hyper-parameter tuning, using a larger dataset with at least a couple thousand images in total (compared to the current dataset with 1998 images) and studying its reconstruction and predictive performance in a full cross validation study where the network is trained on a train/validation/test split dataset. These comments have been discussed in more depth in the final section of Chapter 7.

Appendix A

Appendix

A.1 Supplementary Figures and Tables

Noise	**Standard Deviation of VE**					
	Proc		**Cov**		**FPCA**	
n=50	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$
0%	0.0406	10^{-15}	0.0406	10^{-15}	0.0408	10^{-17}
10%	0.0414	0.0173	0.0414	0.0173	0.0412	0.0175
20%	0.0381	0.0263	0.0381	0.0263	0.0380	0.0264
n=100						
0%	0.0339	10^{-15}	0.0339	10^{-15}	0.0341	0
10%	0.0299	0.0141	0.0299	0.0140	0.0300	0.0142
20%	0.0283	0.0207	0.0282	0.02065	0.0283	0.0208
n=1000						
0%	0.0097	10^{-15}	0.0097	10^{-15}	0.0097	0
10%	0.0093	0.0044	0.0093	0.0044	0.0093	0.0044
20%	0.0087	0.0072	0.0087	0.0072	0.0087	0.0072

Table A.1: D1.1 result: standard deviation of VE per PC (cumulative) over 100 replicates.

Noise	**Standard Deviation of VE**										
	Proc			**Cov**					**FPCA**		
n=50	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_3(s)$	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_3(s)$	$\hat{\phi}_4(s)$	$\hat{\phi}_5(s)$	$\hat{\phi}_1(s)$	$\hat{\phi}_2(s)$	$\hat{\phi}_3(s)$
0%	0.0185	0.0308	0.0136	0.0343	0.0243	0.0129	0.0024	0.0016	0.0464	0.0373	0.0000
10%	0.0181	0.0351	0.0143	0.0373	0.0237	0.0139	0.0080	0.0078	0.0410	0.0342	0.0081
20%	0.0174	0.0327	0.0185	0.0348	0.0246	0.0182	0.0165	0.0170	0.0391	0.0344	0.0169
n=100											
0%	0.0149	0.0266	0.0103	0.0286	0.0186	0.0100	0.0018	0.0010	0.0309	0.0265	0.0000
10%	0.0146	0.0279	0.0106	0.0280	0.0174	0.0103	0.0060	0.0055	0.0337	0.0265	0.0060
20%	0.0139	0.0264	0.0134	0.0264	0.0177	0.0132	0.0120	0.0115	0.0321	0.0266	0.0124
n=1000											
0%	0.0036	0.0085	0.0035	0.0076	0.0063	0.0033	0.0006	0.0004	0.0100	0.0083	0.0000
10%	0.0036	0.0082	0.0037	0.0074	0.0061	0.0035	0.0018	0.0017	0.0098	0.0081	0.0018
20%	0.0035	0.0077	0.0048	0.0070	0.0065	0.0047	0.0039	0.0039	0.0094	0.0080	0.0040

Table A.2: D2.1 result: average VE per PC (cumulative) over 100 replicates.

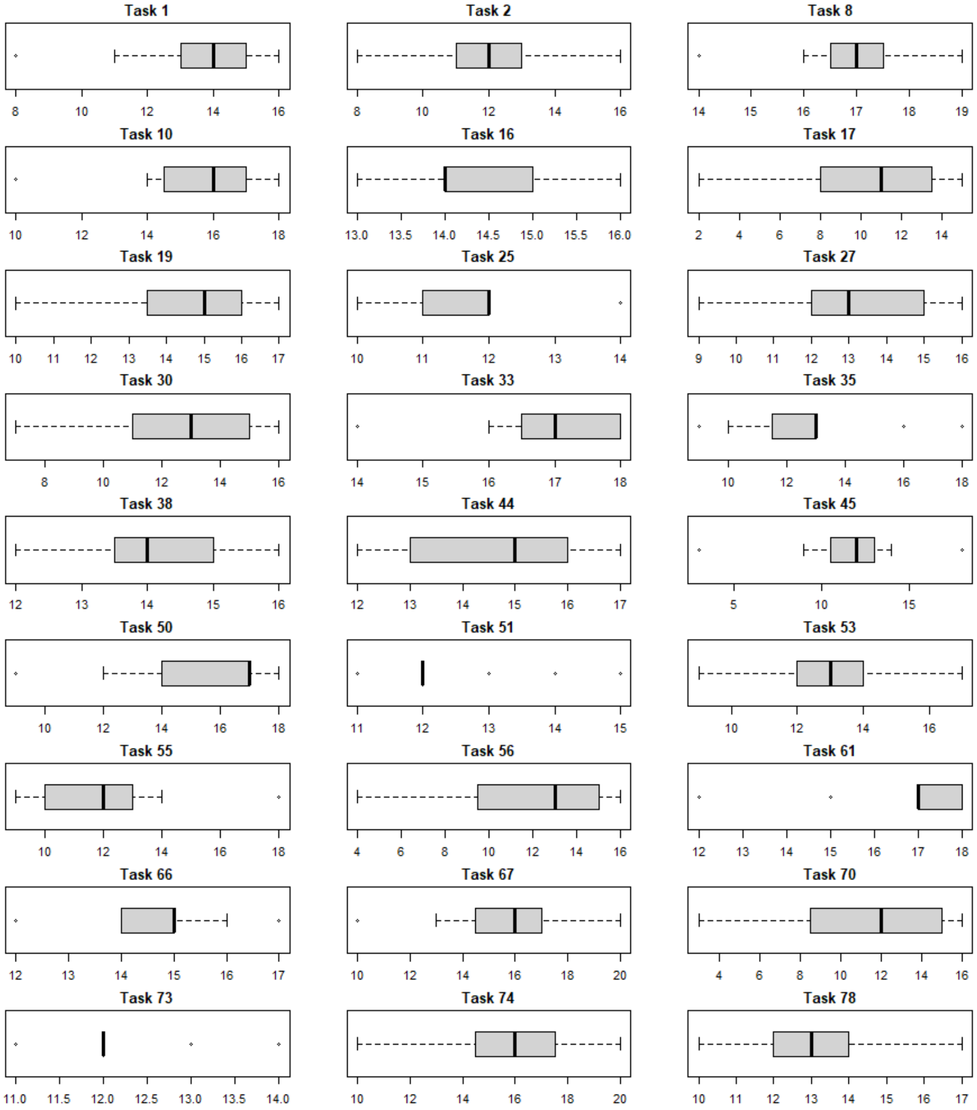


Figure A.1: Subject response prevalence in tasks of type 3 presented in box-plot form. The responses are pseudo-continuous on a scale from 0 to 21.

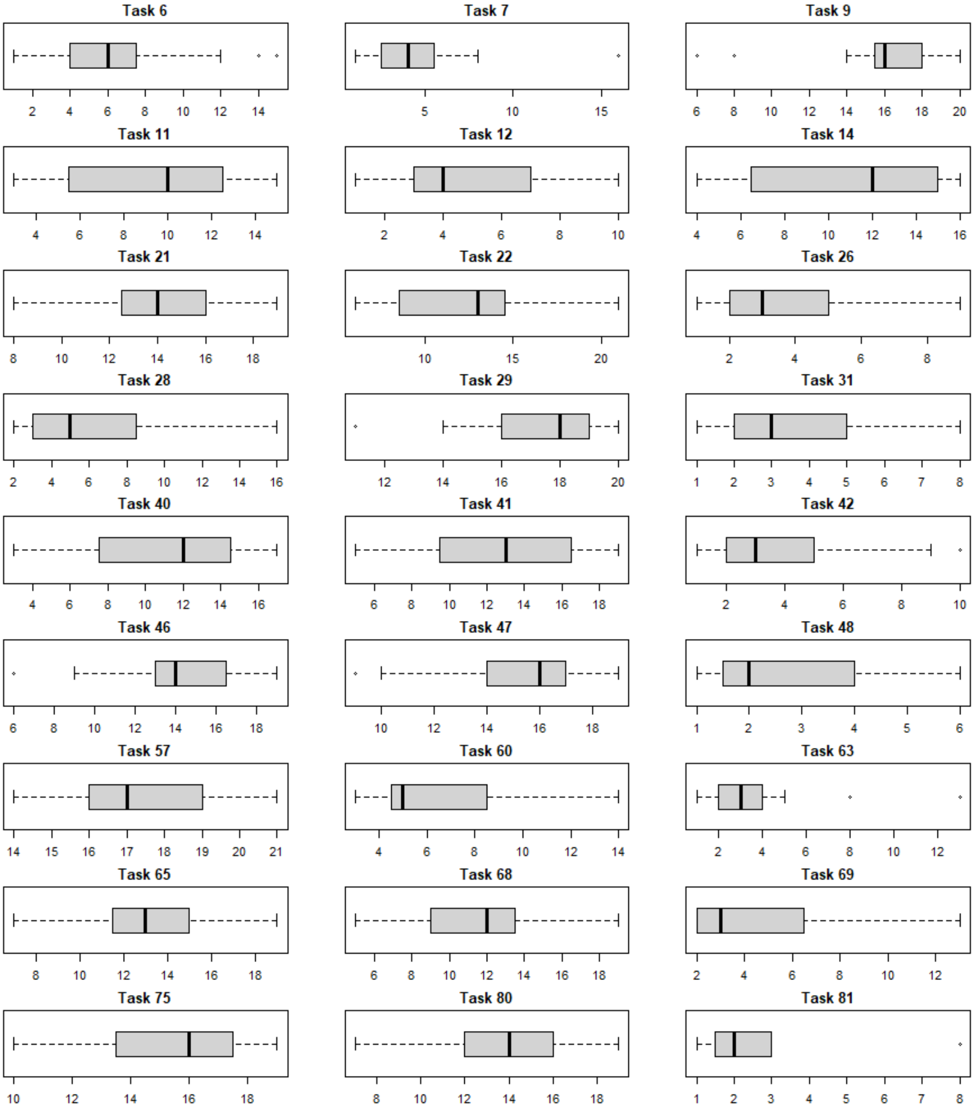


Figure A.2: Subject response prevalence in tasks of type 4 presented in box-plot form. The responses are pseudo-continuous on a scale from 0 to 21.

Bibliography

Aizerman, M. A., Braverman, E. A., & Rozonoer, L. (1964). Theoretical Foundations of the Potential Function method in Pattern Recognition Learning. *Automation and Remote Control, 25*(6), 821–837 (cit. on p. 10).

Almli, C. R., Rivkin, M. J., & McKinstry, R. C. (2007). The NIH MRI study of normal brain development (Objective-2): Newborns, infants, toddlers, and preschoolers. *NeuroImage*, *35*(1), 308–325. https://doi.org/10.1016/j.neuroimage.2006.08.058 (cit. on p. 31).

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronné, R., Faouzi, J., Koval, I., Louis, M., Thibeau-Sutre, E., Wen, J., Wild, A., Burgos, N., Dormont, D., Colliot, O., & Durrleman, S. (2021). Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. *Medical Image Analysis*, *67*, 101848. https://doi.org/10.1016/j.media.2020.101848 (cit. on p. 41).

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry - The methods. *NeuroImage*, *11*(6 I), 805–821. https://doi.org/10.1006/nimg.2000.0582 (cit. on p. 30).

Ashton, E., & Du, T. (2004). Semi-automated measurement of anatomical structures using statistical and morphological priors. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*. https://doi.org/10.1117/12.533047 (cit. on p. 28).

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004 (cit. on p. 42).

Avants, B., Tustison, N., & Johnson, H. (2009). Advanced Normalization Tools (ANTS). *Insight Journal*, 1–35 (cit. on p. 29).

Avants, B. B., Tustison, N. J., Stauffer, M., Song, G., Wu, B., & Gee, J. C. (2014). The Insight ToolKit image registration framework. *Frontiers in Neuroinformatics*, *8*(44), 1–13. https://doi.org/10.3389/fninf.2014.00044 (cit. on p. 42).

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01 (cit. on p. 97).

Besse, P., & Ramsay, J. O. (1986). Principal Components Analysis of Sample Functions. *Psychometrika*, *51*(2), 285–311 (cit. on p. 22).

Bosq, D. (2000). Lieanr Process in Function Space. *Springer Science and Business Media* (cit. on pp. 15, 22).

Capuano, A. W., Wilson, R. S., Leurgans, S. E., Dawson, J. D., Bennett, D. A., & Hedeker, D. (2018). Sigmoidal mixed models for longitudinal data. *Statistical Methods in Medical Research*, *27*(3), 863–875. https://doi.org/10.1177/0962280216645632 (cit. on p. 41).

Carass, A., Cuzzocreo, J., Wheeler, M. B., Bazin, P. L., Resnick, S. M., & Prince, J. L. (2011). Simple paradigm for extra-cerebral tissue removal: Algorithm and analysis. *NeuroImage*, *56*(4), 1982–1992. https://doi.org/10.1016/j.neuroimage.2011.03.045 (cit. on p. 33).

Castro, P. E., Lawton, W. H., & Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics*, *28*(4), 329–337. https://doi.org/10.1080/00401706.1986.10488151 (cit. on p. 22).

Chen, K., & Müller, H. G. (2012). Modeling repeated functional observations. *Journal of the American Statistical Association*, *107*(500), 1599–1609. https://doi.org/10.1080/01621459.2012.734196 (cit. on pp. 3, 25, 44, 75, 94).

Chen, X. R., Shao, Y., & Sadowski, M. J. (2021). Segmented linear mixed model analysis reveals association of the APOE 4 allele with faster rate of alzheimer's disease dementia progression. *Journal of Alzheimer's Disease*, *82*(3), 921–937. https://doi.org/10.3233/JAD-210434 (cit. on p. 40).

Chen, Y., Härdle, W. K., Qiang, H., & Majer, P. (2015). Risk related brain regions detected with 3D image FPCA. *Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin This* (cit. on pp. 37, 44, 65, 76).

Chen, Z., Yi, G. Y., & Wu, C. (2014). Marginal analysis of longitudinal ordinal data with misclassification in both response and covariates. *Biometrical Journal*, *56*(1), 69–85. https://doi.org/10.1002/bimj.201200195 (cit. on p. 72).

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395 (cit. on p. 32).

Dauxois, J., Pousse, A., & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, *12*(1), 136–154. https://doi.org/10.1016/0047-259X(82)90088-4 (cit. on pp. 15, 22).

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–102. https://doi.org/10.1214/ss/1038425655 (cit. on pp. 17, 53).

Elnakib, A. (2013). Developing Advanced Mathematical Models for Detecting Abnormalities in 2D/3D Medical Structures. *Doctoral Dissertation, MAnsoura University* (cit. on p. 27).

Evans, A. C. (2006). The NIH MRI study of normal brain development. *NeuroImage*, *30*(1), 184–202. https://doi.org/10.1016/j.neuroimage.2005.09.068 (cit. on p. 31).

Ferraty, F., & Romain, Y. (2011). *The Oxford Handbook of Functional Data Analysis* (Vol. 67). Oxford University Press. https://doi.org/10.1111/j.1541-0420.2011.01704.x (cit. on pp. 1, 7).

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033 (cit. on pp. 31, 42).

Fouladi, S., Safaei, A. A., Arshad, N. I., Ebadi, M. J., & Ahmadian, A. (2022). The use of artificial neural networks to diagnose Alzheimer's disease from brain images. *Multimedia Tools and Applications*, *81*(26), 37681–37721. https://doi.org/10.1007/s11042-022-13506-7 (cit. on p. 41).

Fox, N. C., Freeborough, P. A., & Rossor, M. N. (1996). Visualisation and quantification of rates of atrophy in Alzheimer's disease. *The Lancet*, *348*, 94–97 (cit. on p. 39).

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *JSS Journal of Statistical Software*, *33*(1) (cit. on p. 97).

Friston, K. J., Josephs, O., Zarahn, E., Holmes, A. P., Rouquette, S., & Poline, J. B. (2000). To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *NeuroImage*, *12*(2), 196–208. https://doi.org/10.1006/nimg.2000.0609 (cit. on p. 33).

Friston, K., Holmes, A., Worsley, K., Poline, J.-p., Frith, C., & Frackowiak, R. (1995). Statistical Parametric Maps in Functional Imaging: A General Linear Approach. *Human Brain Mapping*, *2*, 189 (cit. on p. 43).

García-Martín, E., Rodrigues, C. F., Riley, G., & Grahn, H. (2019). Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, *134*, 75–88. https://doi.org/10.1016/j.jpdc.2019.07.007 (cit. on p. 125).

Georgiou, S., Kechagia, M., Sharman, T., Sarro, F., & Zou, Y. (2022). Green AI: Do Deep Learning Frameworks Have Different Costs? *International Conference on Software Engineering (ICSE)* (cit. on p. 125).

Gerstorf, D., Ram, N., Mayraz, G., Hidajat, M., Lindenberger, U., Wagner, G. G., & Schupp, J. (2010). Late-life decline in well-being across adulthood in germany, the united kingdom, and the united states: Something is seriously wrong at the end of life. *Psychology and Aging*, *25*(2), 477–485. https://doi.org/10.1037/a0017543 (cit. on p. 41).

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). Reproducing Kernel Hilbert Spaces, Mercer's Theorem, Eigenfunctions, Nustrom Method, and Use of Kernels in Machine Learning: Tutorial and Survey (cit. on p. 10).

Greven, S., Crainiceanu, C., Caffo, B., & Reich, D. (2010). Longitudinal Functional Principal Component Analysis. *Electronic Journal of Statistics*, *4*, 1022–1054. https://doi.org/10.1214/10-EJS575 (cit. on pp. 3, 25, 75, 94).

Gromenko, O., & Kokoszka, P. (2013). Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination. *Computational Statistics and Data Analysis*, *59*(1), 82–94. https://doi.org/10.1016/j.csda.2012.09.016 (cit. on pp. 25, 94).

Gromenko, O., Kokoszka, P., Zhu, L., & Sojka, J. (2012). Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *Annals of Applied Statistics*, *6*(2), 669–696. https://doi.org/10.1214/11-AOAS524 (cit. on pp. 25, 94).

Grueso, S., & Viejo-Sobera, R. (2021). Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. *Alzheimer's Research and Therapy*, *13*(1). https://doi.org/10.1186/s13195-021-00900-w (cit. on p. 41).

Guerrero, R., Schmidt-Richberg, A., Ledig, C., Tong, T., Wolz, R., Rueckert, D., Alzheimer's Disease, T., Initiative, N., & Guerrero, R. (2016). Instantiated mixed effects modeling of Alzheimer's disease markers. *NeuroImage*, *142*, 113–125 (cit. on pp. 41, 95).

Guo, W. (2002). Functional Mixed Effects Models. *Biometrics*, *58*, 121–128 (cit. on p. 41).

Hall, C. B., Lipton, R. B., Sliwinski, M., & Stewart, W. F. (2000). A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. *Statistics in Medicine*, *19*(11-12), 1555–1566. https://doi.org/10.1002/(SICI)1097-0258(20000615/30)19:11/12¡1555::AID-SIM445¿3.0.CO;2-3 (cit. on p. 41).

Happ, C., & Greven, S. (2018). Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, *113*(522), 649–659. https://doi.org/10.1080/01621459.2016.1273115 (cit. on p. 41).

Horvath, L., & Kokoszka, P. (2012). Inference for Functional Data with Applications. *Springer Series in Statistics* (cit. on pp. 1, 7, 15, 21).

Hsing, T., & Eubank, R. (2015). Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. *Wiley & Sons* (cit. on p. 7).

Jack, C. R., Petersen, R. C., O'Brien, P. C., & Tangalos, E. G. (1992). MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. *Neurology*, *42*(1), 183–183. https://doi.org/10.1212/WNL.42.1.183 (cit. on p. 39).

Jack, C. R., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Lowe, V., Kantarci, K., Bernstein, M. A., Senjem, M. L., Gunter, J. L., Boeve, B. F., Trojanowski, J. Q., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., & Knopman, D. S. (2012). Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Archives of Neurology*, *69*(7), 856–867. https://doi.org/10.1001/archneurol.2011.3405 (cit. on p. 40).

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015 (cit. on p. 29).

Juntu, J., Sijbers, J., Dyck, D., & Gielen, J. (2008). Bias Field Correction for MRI Images. (January), 543–551. https://doi.org/10.1007/3-540-32390-2–“˙”64 (cit. on p. 30).

Kassinopoulos, M., & Mitsis, G. D. (2022). A multi-measure approach for assessing the performance of fMRI preprocessing strategies in resting-state functional connectivity. *Magnetic Resonance Imaging*, *85*(March 2021), 228–250. https://doi.org/10.1016/j.mri.2021.10.028 (cit. on p. 34).

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes (cit. on pp. 111–113).

Lancaster, J. L., Tordesillas-Gutiérrez, D., Martinez, M., Salinas, F., Evans, A., Zilles, K., Mazziotta, J. C., & Fox, P. T. (2007). Bias between MNI and talairach coordinates analyzed using the ICBM-152 brain template. *Human Brain Mapping*, *28*(11), 1194–1205. https://doi.org/10.1002/hbm.20345 (cit. on p. 31).

Lemieux, L., Hagemann, G., Krakow, K., & Woermann, F. G. (1999). Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. *Magnetic Resonance in Medicine*, *42*(1), 127–135. https://doi.org/10.1002/(SICI)1522-2594(199907)42:1¡127::AID-MRM17¿3.0.CO;2-O (cit. on p. 32).

Leong, R. L., Lo, J. C., Sim, S. K., Zheng, H., Tandi, J., Zhou, J., & Chee, M. W. (2017). Longitudinal brain structure and cognitive changes over 8 years in an East Asian cohort. *NeuroImage*, *147*, 852–860. https://doi.org/10.1016/j.neuroimage.2016.10.016 (cit. on p. 39).

Li, C., Xiao, L., & Luo, S. (2022). Joint model for survival and multivariate sparse functional data with application to a study of Alzheimer's Disease. *Biometrics*, *78*(2), 435–447. https://doi.org/10.1111/biom.13427 (cit. on p. 41).

Li, D., Iddi, S., Thompson, W. K., Rafii, M. S., Aisen, P. S., & Donohue, M. C. (2018). Bayesian latent time joint mixed-effects model of progression in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *10*, 657–668. https://doi.org/10.1016/j.dadm.2018.07.008 (cit. on p. 41).

Li, Y., Huang, C., & Härdle, W. K. (2019). Spatial functional principal component analysis with applications to brain image data. *Journal of Multivariate Analysis*, *170*, 263–274. https://doi.org/10.1016/j.jmva.2018.11.004 (cit. on pp. 24, 37, 38, 43–45, 64, 67, 70–72, 127).

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., & Lancaster, J. (1995). A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). *NeuroImage*, *2*(2), 89–101 (cit. on p. 31).

Mckeown, M. J., Jung, T.-P., Makeig, S., Brown, G., Kindermann, S. S., Lee, T.-W., & Sejnowski, T. J. (1998). Neuroimaging of Human Brain Function. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 803–810 (cit. on p. 44).

Mercer, J. (1909). Functions of positive and negative type, and their connection the theory of integral equations. *Proceedings of the Royal Society of London. Series A, Containing*

Papers of a Mathematical and Physical Character, *209*(441-458), 415–446. https://doi.org/https://doi.org/10.1098/rsta.1909.0016 (cit. on p. 10).

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., . . . Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393 (cit. on p. 95).

Milliken, J. K., & Edland, S. D. (2000). Mixed effect models of longitudinal Alzheimer's disease data: A cautionary note. *Statistics in Medicine*, *19*(11-12), 1617–1629. https://doi.org/10.1002/(SICI)1097-0258(20000615/30)19:11/12¡1617::AID-SIM450¿3.0.CO;2-C (cit. on pp. 41, 108).

Mofrad, S. A., Lundervold, A., & Lundervold, A. S. (2021). A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, *90*. https://doi.org/10.1016/j.compmedimag.2021.101910 (cit. on pp. 41, 95).

Mohr, P. N., Biele, G., & Heekeren, H. R. (2010a). Neural processing of risk. *Journal of Neuroscience*, *30*(19), 6613–6619. https://doi.org/10.1523/JNEUROSCI.0003-10.2010 (cit. on pp. 2, 35, 36).

Mohr, P. N., Biele, G., Krugel, L. K., Li, S. C., & Heekeren, H. R. (2010b). Neural foundations of risk–return trade-off in investment decisions. *NeuroImage*, *49*(3), 2556–2563. https://doi.org/10.1016/J.NEUROIMAGE.2009.10.060 (cit. on pp. 34, 37, 65, 66).

Monti, M. M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Frontiers in Human Neuroscience*, *5*(1). https://doi.org/10.3389/fnhum.2011.00028 (cit. on p. 43).

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

Dementia, *1*(1), 55–66. https://doi.org/10.1016/j.jalz.2005.06.003 (cit. on pp. 2, 38, 95).

Narai, A., Hermann, P., Kemenczky, P., Szalma, J., Homolya, I., Somogyi, E., Vakli, P., Weiss, B., & Vidnyanszky, Z. (2022). Movement-related artefacts (MR-ART) dataset of matched motion-corrupted and clean structural MRI brain scans. *Scientific Data*, *9*(630). https://doi.org/https://doi.org/10.1038/s41597-022-01694-8 (cit. on p. 29).

Nyul, L., Udupa, J. K., & Zhang, X. (2000). New Variants of a Method of MRI Scale Standardization. *IEEE Transactions on Medical Imaging*, *19*(2), 143–150 (cit. on p. 32).

Palma, M., Tavakoli, S., Brettschneider, J., & Nichols, T. E. (2020). Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression. *NeuroImage*, *219*, 116938. https://doi.org/10.1016/J.NEUROIMAGE.2020.116938 (cit. on p. 42).

Park, A. Y., Aston, J. A. D., & Ferraty, F. (2016). Stable and predictive functional domain selection with application to brain images (cit. on p. 23).

Park, B. U., Mammen, E., Härdle, W., & Borak, S. (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association*, *104*(485), 284–298. https://doi.org/10.1198/jasa.2009.0105 (cit. on pp. 3, 44).

Park, D. C., & Reuter-Lorenz, P. (2009). The Adaptive Brain: Aging and Neurocognitive Scaffolding. *Annual Review of Psychology*, *60*(1), 173–196. https://doi.org/10.1146/annurev.psych.59.103006.093656 (cit. on p. 39).

Park, S. Y., & Staicu, A. M. (2015). Longitudinal functional data analysis. *Stat*, *4*(1), 212–226. https://doi.org/10.1002/sta4.89 (cit. on pp. 3, 4, 25, 43–45, 75, 76, 78, 83, 92–94, 129).

Petersen, K., Hansen, L., Kolenda, T., Rostrup, E., & Strother, S. C. (2000). On the Independent Components of Functional Neuroimages. *Third International Conference on Independent Component Analysis and Blind Source Separation*, 615–620 (cit. on p. 44).

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Weiner, M. W.

(2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, *74*(3), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25 (cit. on p. 38).

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). Handbook of Functional MRI Data Analysis. *Cambridge University Press* (cit. on p. 26).

Polzehl, J., Papafitsoros, K., & Tabelow, K. (2020). Patch-wise adaptive weights smoothing in r. *Journal of Statistical Software*, *95*, 1–27. https://doi.org/10.18637/jss.v095.i06 (cit. on p. 102).

Preda, C., & Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, *48*(1), 149–158. https://doi.org/10.1016/j.csda.2003.10.003 (cit. on p. 22).

Preda, C., Saporta, G., & Lévéder, C. (2007). PLS classification of functional data. *Computational Statistics*, *22*(2), 223–235. https://doi.org/10.1007/s00180-007-0041-4 (cit. on p. 22).

Prince, J. L., & Links, J. (2015). Medical Imaging: Signals and Systems. *Pearson Prentice Hall*, 1–10 (cit. on p. 26).

Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, *47*(4), 379–396 (cit. on p. 1).

Ramsay, J. O., & Silverman, B. W. (2005). Functional Data Analysis. *Springer*, 426 (cit. on pp. 1, 7, 16–18, 21, 22, 44, 129).

Reinhold, J. C., Dewey, B. E., Carass, A., & Prince, J. L. (2019). Evaluating the impact of intensity normalization on MR image synthesis, 126. https://doi.org/10.1117/12.2513089 (cit. on p. 32).

Rice, J., & Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(1), 233–243 (cit. on p. 22).

Riesz, F., & Sz.-Nagy, B. (1956). Functional Analysis (L. Boron, Ed.; 2nd). *Blackie & Son Limited* (cit. on p. 10).

Routier, A., Burgos, N., Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Hassanaly, R., Jacquemont, T., Lu, P., Marcoux, A., Moreau, T., Samper-González, J., Teichmann, M., Thibeau-Sutre, E., Vaillant, G., . . . Colliot, O. (2021). Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies. *Frontiers in Neuroinformatics*, *15*, 1–16. https://doi.org/10.3389/fninf.2021.689675 (cit. on p. 42).

Rowe, T. W., Katzourou, I. K., Stevenson-Hoare, J. O., Bracher-Smith, M. R., Ivanov, D. K., & Escott-Price, V. (2021). Machine learning for the life-time risk prediction of Alzheimer's disease: a systematic review. *Brain Communications*, *3*(4). https://doi.org/10.1093/braincomms/fcab246 (cit. on p. 41).

Sauty, B., & Durrleman, S. (2022). Progression models for imaging data with Longitudinal Variational Auto Encoders Progression models for imaging data with Longitudinal Variational Auto Encoders Progression models for imaging data with Longitudinal Variational Auto Encoders. *International Conference on Medical Image Computing and Computer Assisted Intervention* (cit. on pp. 110–112).

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., & Arbel, T. (2011). Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis*, *15*(2), 267–282. https://doi.org/10.1016/j.media.2010.12.003 (cit. on p. 32).

Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., & Leahy, R. M. (2001). Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, *13*(5), 856–876. https://doi.org/10.1006/nimg.2000.0730 (cit. on p. 33).

Shi, H., Ma, D., Nie, Y., Faisal Beg, M., Pei, J., Cao, J., & Neuroimaging Initiative, T. A. D. (2021). Early diagnosis of Alzheimer's disease on ADNI data using novel longitudinal score based on functional principal component analysis. *Journal of Medical Imaging*, *8*(02), 1–16. https://doi.org/10.1117/1.jmi.8.2.024502 (cit. on p. 41).

Silverman, B. W. (1996). Smoothed Functional Principal Components Analysis by Choice of Norm. *The Annals of Statistics*, *24*(1), 1–24 (cit. on p. 22).

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, *17*(1), 87–97. https://doi.org/10.1109/42.668698 (cit. on pp. 30, 31).

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062 (cit. on p. 33).

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*(SUPPL. 1), 208–219. https://doi.org/10.1016/j.neuroimage.2004.07.051 (cit. on p. 29).

Stella Atkins, M., & Mackiewich, B. T. (1998). Fully automatic segmentation of the brain in MRI. *IEEE Transactions on Medical Imaging*, *17*(1), 98–107. https://doi.org/10.1109/42.668699 (cit. on p. 32).

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for modern deep learning research. *arXiv*, (1), 1393–13696. https://doi.org/10.1609/aaai.v34i09.7123 (cit. on p. 125).

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). 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), 1–10. https://doi.org/10.1371/journal.pmed.1001779 (cit. on p. 95).

Thompson, P. M., Hayashi, K. M., De Zubicaray, G. I., Janke, A. L., Rose, S. E., Semple, J., Hong, M. S., Herman, D. H., Gravano, D., Doddrell, D. M., & Toga, A. W. (2004).

Mapping hippocampal and ventricular change in Alzheimer disease. *NeuroImage*, *22*(4), 1754–1766. https://doi.org/10.1016/j.neuroimage.2004.03.040 (cit. on pp. 39, 108).

Thompson, W. K., & Rosen, O. (2008). A Bayesian model for sparse functional data. *Biometrics*, *64*(1), 54–63. https://doi.org/10.1111/j.1541-0420.2007.00829.x (cit. on p. 41).

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288 (cit. on p. 97).

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of Neurophysiology*, *97*(2), 1621–1632. https://doi.org/10.1152/jn.00745.2006 (cit. on pp. 37, 65).

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908 (cit. on p. 42).

Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., Kandel, B. M., van Strien, N., Stone, J. R., Gee, J. C., & Avants, B. B. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *NeuroImage*, *99*, 166–179. https://doi.org/10.1016/j.neuroimage.2014.05.044 (cit. on p. 30).

van Bömmel, A., Song, S., Majer, P., Mohr, P. N., Heekeren, H. R., & Härdle, W. K. (2013). Risk Patterns and Correlated Brain Activities. Multidimensional Statistical Analysis of fMRI Data in Economic Decision Making Study. *Psychometrika*, *79*(3), 489–514. https://doi.org/10.1007/s11336-013-9352-2 (cit. on pp. 3, 37, 38, 44, 66).

van Dijk, K. R., Sabuncu, M. R., & Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage*, *59*(1), 431–438. https://doi.org/10.1016/j.neuroimage.2011.07.044 (cit. on p. 34).

Veitch, D. P., Weiner, M. W., Miller, M., Aisen, P. S., Ashford, M. A., Beckett, L. A., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Landau, S. M., Morris, J. C., Nho, K. T., Nosheny, R., Okonkwo, O., Perrin, R. J., Petersen, R. C., Rivera Mindt, M., Saykin, A., . . . Tosun, D. (2024). The Alzheimer's Disease Neuroimaging Initiative in the era of Alzheimer's disease treatment: A review of ADNI studies from 2021 to 2022. *Alzheimer's and Dementia*, *20*(1), 652–694. https://doi.org/10.1002/alz.13449 (cit. on p. 39).

Wang, J.-L., Chiou, J.-M., & Mueller, H.-G. (2015). Review of Functional Data Analysis, 1–41. https://doi.org/10.1146/)) (cit. on pp. 2, 23, 44).

Wang, X., Nan, B., Zhu, J., & Koeppe, R. (2014a). Reularised 3D Functional Regression for Brain Image Data via Haar Wavelets. *Annals of Applied Statistics*, *8*(2). https://doi.org/10.1214/14-AOAS736.REGULARIZED (cit. on p. 23).

Wang, X., Nan, B., Zhu, J., Koeppe, R., & Frey, K. (2017). Classification of ADNI PET images via regularized 3D functional data analysis. *Biostatistics and Epidemiology*, *1*(1), 3–19. https://doi.org/10.1080/24709360.2017.1280213 (cit. on pp. 23, 42).

Wang, Y., Nie, J., Yap, P. T., Li, G., Shi, F., Geng, X., Guo, L., & Shen, D. (2014b). Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. *PLoS ONE*, *9*(1), 1–23. https://doi.org/10.1371/journal.pone.0077810 (cit. on p. 33).

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L. M., Toga, A. W., & Trojanowski, J. Q. (2017). Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's and Dementia*, *13*(4), e1–e85. https://doi.org/10.1016/j.jalz.2016.11.007 (cit. on p. 39).

Weiner, M. W., Veitch, D. P., Miller, M. J., Aisen, P. S., Albala, B., Beckett, L. A., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Landau, S. M., Morris, J. C., Nosheny, R., Okonkwo, O. C., Perrin, R. J., Petersen, R. C., Rivera-Mindt, M., Saykin, A. J., Shaw, L. M., . . . Trojanowski, J. Q. (2023). Increasing participant diversity in AD research: Plans for digital screening, blood testing, and a community-engaged approach in the

Alzheimer's Disease Neuroimaging Initiative 4. *Alzheimer's and Dementia*, *19*(1), 307–317. https://doi.org/10.1002/alz.12797 (cit. on p. 39).

Weiner, M., & ADNI. (2013). The ADNI initiative: review of paper published since its inception. *Alzheimer Dementia*, *9*(5), e111–e194. https://doi.org/10.1016/j.jalz.2013.05.1769.The (cit. on p. 39).

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., & Colliot, O. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, *63*, 101694. https://doi.org/10.1016/j.media.2020.101694 (cit. on p. 42).

West, M. J., Coleman, P. D., Flood, D. G., & Troncoso, J. C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *The Lancet*, *344*(8925), 769–772 (cit. on p. 39).

Whitcher, B., Schmid, V., & Thornton, A. (2011). Working with the DICOM and NIfTI Data Standards in R. *Journal of Statistical Software*, *44*(6), 1–28. https://doi.org/10.18637/jss.v044.i06 (cit. on p. 102).

Wyman, B. T., Harvey, D. J., Crawford, K., Bernstein, M. A., Carmichael, O., Cole, P. E., Crane, P. K., Decarli, C., Fox, N. C., Gunter, J. L., Hill, D., Killiany, R. J., Pachai, C., Schwarz, A. J., Schuff, N., Senjem, M. L., Suhy, J., Thompson, P. M., Weiner, M., & Jack, C. R. (2013). Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's and Dementia*, *9*(3), 332–337. https://doi.org/10.1016/j.jalz.2012.06.004 (cit. on p. 39).

Xiao, L., Li, Y., Ruppert, D., Schultz, A., & of Engineering, P. (2012). Fast Bivariate P-splines: the Sandwich Smoother (cit. on pp. 52, 53).

Yao, F., Muller, H. G., & Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, *33*(6), 2873–2903. https://doi.org/10.1214/009053605000000660 (cit. on pp. 22, 41).

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., & Crainiceanu, C. (2011a). Funcitonal principal component model for high-dimensional brain imaging. *Neuroimagw*, *3*(58), 772–784. https://doi.org/10.1016/j.neuroimage.2011.05.085 (cit. on pp. 24, 44).

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., & Crainiceanu, C. (2011b). Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, *20*(4), 852–873. https://doi.org/10.1198/jcgs.2011.10122 (cit. on p. 24).

Zou, H., Xiao, L., Zeng, D., & Luo, S. (2023). Multivariate functional mixed model with MRI data: An application to Alzheimer's disease. *Statistics in Medicine*, *42*(10), 1492–1511. https://doi.org/10.1002/sim.9683 (cit. on p. 41).