Statistical learning with elastic functional data analysis

by

Mengyun Yu

(Under the Direction of Cheolwoo Park and Justin Strait)

Abstract

We propose new methods inspired by elastic data analysis to spatially-dependent functional data. We first propose a new classification algorithm of fMRI data based on the notion of amplitude and phase variability separation and a new between-voxel similarity measure based on the spatial dependence structure of the fMRI data. The approach is based on support vector machine (SVM) with a new kernel. The new kernel varies by the functional structure of interest in the brain, and it also captures the amplitude variability of the activation in the region of interest (ROI). This classification algorithm is illustrated via simulated datasets as well as a real dataset on schizophrenia, and its performance is compared with existing SVM based algorithms.

Second, we propose a new Bayesian registration model that incorporates the spatial dependence structure in the priors. We propose a Bayesian registration model with a geometrically-inspired conditional autoregressive prior on warping functions. The CAR prior

allows for warping functions to be dependent a priori, with dependencies characterized by a graph representing spatial structure of the data. The posterior is sampled via a MCMC algorithm that is robust to the discretization of the observed functions. This graph-based Bayesian registration algorithm is shown to be robust to outliers in a simulation study. Its performance is compared with an existing Bayesian registration algorithm.

INDEX WORDS: Bayesian hierarchical models, Classification, Elastic functional data analysis, Spatial temporal data analysis.

Statistical learning with elastic functional data analysis

by

Mengyun Yu

B.S., Fudan Univesity, China, 2016

A Dissertation Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2022

©2022

Mengyun Yu

All Rights Reserved

Statistical learning with elastic functional data analysis

by

Mengyun Yu

Major Professor: Cheolwoo Park

Justin Strait

Committee: Jeongyoun Ahn

Yuan Ke

Nicole Lazar

Electronic Version Approved:

Ron Walcott

Vice Provost for Graduate Education and Dean of the Graduate School

The University of Georgia

August 2022

Acknowledgments

This work was made possible through the support of many people. Words cannot express my gratitude to my major professors Dr. Cheolwoo Park and Dr. Justin Strait for their invaluable guidance and support on each step of the journey. I am also thankful to my defense committee, Dr. Jeongyoun Ahn, Dr. Yuan Ke, and Dr. Nicole Lazar, who generously provided knowledge and expertise. Additional thanks to the UGA neuroimaging group, especially Brooke Jackson, who organized and provided the data used in this work, for the great opportunities to share and learn.

I am grateful to the staff and faculties that helped me transition smoothly into a foreign country for study and encouraged me to pursue a Ph.D. degree. I also want to thank my supervisors during my assistantships, Dr. Ford Ballantyne, Dr. Dan Hall, and Dr. Xianyan Chen for their support and recognition of my work. Special thanks to Dr. Cheolwoo Park and Dr. Jaxk Reeves for their guidance in the early time and towards the end of the program.

Lastly, I would be remiss in not mentioning my family, especially my parents. Their belief in me has kept my spirits and motivation high during this process.

Contents

Acknowledgments **iv**

List of Figures **vi**

List of Tables **viii**

1 Introduction **1**

1.1 Function registration 1

1.2 Tasks of interest 6

2 Classification of fMRI data **8**

2.1 FMRI data 9

2.2 Elastic functional data analysis 14

2.3 Support vector machine (SVM) 21

2.4 Proposed approach for fMRI classification 26

2.5 Simulation and real data applications 31

2.6 Discussion 44

3 Bayesian framework on the spatial temporal data **47**

3.1 Review of Bayesian registration methods 48
3.2 Bayesian registration via Gaussian process priors 49
3.3 Elastic graph-based function registration 61
3.4 Simulation study 71
3.5 A real data application 76
3.6 Discussion 79

4 Conclusion **81**

Bibliography **83**

List of Figures

2.1 HRF 11

2.2 Illustration of pinching effect using $\mathcal{L}^2$ distance [33] 18

2.3 Illustration of finding the mean template in $[\mu_q]$ with respect to $(\gamma_1, \cdots, \gamma_n)$ 21

2.4 Illustration of the SVM algorithm 22

2.5 Comparison of registration by voxels and by point-wise means 31

2.6 Activation signals from setup I 34

2.7 Point-wise average of the simulated fMRI data I 34

2.8 Registration result on simulated data I 36

2.9 Performance comparison based on the simulated data I 37

2.10 Activation signals 38

2.11 Point-wise average of the simulated fMRI data 39

2.12 Registration of the functions in simulated data II 40

2.13 Performance comparison based on the simulated data II 41

3.1 Illustration of the transformations of the warping space 51

3.2 Registration results for the Berkeley Growth Study [23] 59

3.3 An example of over-fitting for the independent Bayesian registration method 61

3.4 The proposed graph structure on the simulated data 72

3.5 Trace plot of the simulation study 73

3.6 A simulation study of the elastic graph-based registration 74

3.7 Registration comparison on the sine wave with higher frequency 75

3.8 Comparison of the registered mean 75

3.9 An application of the graph-based registration on the fMRI study 77

3.10 Bayesian registration comparison on the warpings of the fMRI data 78

3.11 Bayesian registration comparison on the function means of the fMRI data . . 78

List of Tables

2.1 Simulation setup I 33

2.2 Hyper parameters for the simulation study 35

2.3 Simulation setup II 38

2.4 20 pre-specified ROIs 43

2.5 Hyper parameters for the real data 43

2.6 Algorithm performance on the real data 44

2.7 Mean classification errors on the real data among 10 runs 44

Chapter 1

Introduction

We first introduce the idea of function registration as well as existing methods and then describe the typical data that we are interested in applying registration (Section 1.1). In Section 1.2, we introduce the two tasks that we aim to answer in the following chapters.

1.1 Function registration

1.1.1 Function Data Analysis (FDA)

Function registration is one important research area in functional data analysis (FDA). The term FDA was coined by Ramsay (1982) [28] and has became an important branch of statistics. FDA specifically deals with data in the form of functions, images and shapes, or more general objects that vary over a continuous space. In the FDA framework, the sample element is a random function. For each subject, one or more functions could be observed. The above continuous space could be time, spatial location, probability and others.

Functional data are intrinsically infinite dimensional, which brings challenges for both theory and computation, but also opportunities for research and analysis. One of the most recent review articles of FDA is given by Wang et al. (2016) [42].

The most popular modeling framework for FDA is assuming that functional data are realizations of a stochastic process. For example, for curve data, or real-valued functions on a compact interval like $I = [0, 1]$, can be assumed to be samples from a one-dimensional stochastic process $X(t)$ in the Hilbert space, such as $\mathcal{L}_2(I)$. Specific definition of a $\mathcal{L}_p$ space is given in Section 2.2.1.

In real life, this underlying process is not observed directly, but as discrete data observed on a fixed or random time gird. The time grid could be dense, which could bring in the issue of high-dimensionality, but this abundance in information is useful, and smoothness assumptions (e.g. $\mathcal{L}_2$-continuous) could be imposed to overcome the curse of dimensionality. In addition, the observed discrete data contain random noises. Those random noises are often assumed to be independent across subjects, but could have more complex temporal correlation or spatial correlation structure.

Some main topics in FDA are functional principal components analysis (FPCA), functional regression, and function registration. *FPCA* deals with dimension reduction of the infinite-dimensional functional data to finite-dimensional random vector of scores. Specifically, the underlying random function $X(t)$ is expanded in a data-dependent functional basis, e.g. obtained from singular value decomposition (SVD) of the covariance matrix, and then the expansion is truncated to achieved dimension reduction. It can explain the most variation given a fixed number of expansion components. *Functional regression* is a functional analog of the multivariate linear model, which can have functional response and/or

functional covariates and functional coefficients, where the functional covariates and coefficients are associated with the inner product in the $\mathcal{L}_2$ space, and/or vector covariates and coefficients. Details and application of those topics of FDA can be found in [27]. The main interest of this dissertation is in function registration, which will be introduced in the next section.

1.1.2 Function registration

Function registration, also known as time warping, curve registration, curve alignment, aims to identify and separate amplitude variation and time variation [29]. It is often important when conducting FPCA and functional regression that the functions' salient features are registered or aligned by suitable monotone transformations, and then analyses are carried out on the transformed functional data. One example where function registration is important is the Berkeley Growth study [41]. In this setup, the temporal variation exists due to the difference in children's biological ages, because they reaches growth spurt in puberty differently in time. In the presence of the temporal variation, the point-wise mean of the functions is not the most informative estimate of the growth rate changes over time, as the peaks and valleys are distorted in this mean representation.

For an observed functional data Y_i, the setup of function registration is assuming Y_i is a composite of the amplitude $X_i \overset{iid}{\sim} X$ and a random warping $h_i \overset{iid}{\sim} h$, i.e. $Y_i(t) = X_i(h_i^{-1}(t)), t \in I$.

Below we introduce three methods: dynamic time warping (DTW), landmarks-based registration, and registration using Fisher-Rao metric.

DTW is a technique introduced in [2] in 1959 and it was a popular tool explored in the area of speech recognition [25]. It views functional data as a series of discrete observations and produces discrete matching between the elements of one series to another. That is, the warping h is restricted to be a piece-wise linear function that passes through every observational time point. In a classic *DTW*, the algorithm aims to minimize the $\mathcal{L}_2$ distance between two series, which is a greedy optimization that can lead to the "pinching effect". (See Figure 2.2 in Section 2.2.)

Landmark registration is a traditional method used in shape analysis [17], where special features in the shape are aligned to the mean location and then a smooth transformation from the mean location to the feature location is introduced. This method works well when the feature landmarks are distinctive and well-presented in the observed functions, but not when the landmarks are missing from some samples or when the samples are too noisy burying the information from the landmarks.

Registration using Fisher-Rao metric introduced by Srivastava et al. (2011) [34] is a registration method proposed in a novel geometric framework, also known as elastic functional data analysis, for function and shape analysis which integrates registration in problem into the whole analysis. Instead of working with the standard $\mathcal{L}_2$ metric in function space, they use the Fisher-Rao Riemannian metric that is invariant to warping. The convenience of such metric is that it is the standard $\mathcal{L}_2$ metric on the square-root velocity function (SRVF) representations of the observed function, so that more traditional vector-based statistical analyses are applicable in this setup. It borrows the ideas from statistical modeling on non-linear manifolds and extend those ideas to build FPCA, function regression models under the Fisher-Rao metric. We will introduce the framework of the registration and the SRVF

representation in Section 1.1, and then develop our functional analysis method based on this framework. More geometric properties and applications can be found in [33].

More recently, Bayesian registration has been studied. It has the advantage of having uncertainty of the estimated warpings and mean, and being able to apply prior information in the modeling process. A short review of Bayesian registration can be found in Section 3.1, and we will focus on a specific Bayesian registration method proposed by Lu [22].

In a majority of the methods mentioned above, it is often assumed that the functional data are observed independently. However, more data objects nowadays are complex functions which are often observed with more sophisticated dependence structures, rendering independence assumptions poorly satisfied. One exemplary data type is the spatially-indexed functional data.

Functional neuroimaging is an area that produces such type of data in larges quantities. Techniques such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI), electro-encephalography (EEG), and magneto-encephalography (MEG) acquire real-time observations of the underlying brain processes on different brain regions under certain experimental conditions [37]. In nature, these time trajectories exhibit similarities if they are obtained from neighbouring locations. For fMRI data for example, functional data from the same ROI are associated with a common functional or anatomic feature, so it is only appropriate to assume that those functional data are not independent. A comprehensive overview on statistical methods in fMRI data analysis can be found in [19].

Genomics data generated from the emerging high-throughput technologies are sometimes spatially-indexed functional data as well. For example, in the study of whole-organ histological and genetic maps (WOHGMs), genomic changes such as in molecular quantities are

investigated for subregions that have different histologies within a single tissue map [24]. The data related to those maps are in nature "functions of genomic location for areal regions of the tissue that are expected to exhibit spatial correlation".

Geographic information systems (GIS) is another source of spatially-indexed functional data. For example, the European Climate Assessment and Dataset [16] provides gridded observation data across Europe, including cloudiness, temperature, precipitation, humidity, pressure, snow and sunshine. Those datasets consist of time series (of e.g. temperature) from a set of stations, which are expected to exhibit spatial correlation. There is also potential temporal variability due to altitude differences in the stations that makes it important to include function registration in the analysis.

In this dissertation, we try to tackle two specific analytical tasks of spatially-indexed functional data in the framework of elastic functional data analysis. We briefly introduce those analytical tasks and our solutions to them in the next section.

1.2 Tasks of interest

The first project is motivated by a study of fMRI data among people diagnosed with schizophrenia and people with high and low cognition. The data generated from this study are spatially-index functional data, and the goal is to learn from the two groups of spatially-indexed functional data so that we can classify subjects based on their brain functional performance during an fMRI study. We incorporate the idea of amplitude and phase variability in classifying fMRI data using a spatial functional Support Vector Machine (fSVM) algorithm. Specific methods and results are shown in Chapter 2.

The second project extends data objects from fMRI data to all types of spatially-indexed functional data. The data object is only a set of spatially-indexed functions where their temporal variation is correlated based on the spatial structure. We propose a Bayesian framework for function alignment and registration using priors that captures the spatial dependence across functions, and look into some properties of the proposed model. Specific methods and results are shown in Chapter 3.

Chapter 2

Classification of fMRI data

FMRI is a popular noninvasive technique for observing neural activity in the human brain by observing changes associated with blood flow. FMRI technique nowadays can localize activities within millimeters and within one to two seconds, and a typical fMRI study often scans hundreds of thousands of locations in the brain for hundreds of time points. Hence, data generated from an fMRI experiment are usually of massive volume and are considered highly complex, making it hard to analyze.

A remarkable number of fMRI studies have been dedicated to the neurobehavioral studies in schizophrenia, a chronic brain disorder. A number of cognitive impairments in schizophrenia have been specifically linked to abnormal activity of certain brain areas [11]. Recently, more studies have focused on using machine learning techniques like Support Vector Machine (SVM) [35], to characterize disease related alternation in brain structure and functions, as well as to make diagnostic discrimination. Application of machine learning methods integrated to fMRI experimental data is a valid, inexpensive, and non-invasive way to detect

brain disorder at an early stage, and hence could play a decisive role in diagnosis and prevention [35].

In this chapter, we focus on the task of improving the predictive ability of subjects with schizophrenia, based on the task-based fMRI data. Specifically, we use ideas from manifold learning and functional data analysis (FDA) to define a functional representation of BOLD signals, which utilizes a warping-invariant metric in the space of fMRI functions to perform elastic alignment of the fMRI time series. Then, we develop a spatial functional SVM for fMRI data using the new functional representation.

This chapter is organized as follows. Section 2.1 introduces the target data objects, while Section 2.2 and Section 2.3 introduce the methodology behind the proposed method. Section 2.4 describes the proposed classification algorithm, and Section 2.5 shows the performance of the proposed method via simulation study and application to real fMRI data. Finally, Section 2.6 gives a short discussion on the limitations of the proposed method.

2.1 FMRI data

In this section, we give a brief introduction on the fMRI data via key terminologies. More details of how fMRI works and properties of fMRI data can be found in [10] .

BOLD (Blood Oxygen Level Dependent) signal

Standard fMRI data measures BOLD contrast, which is discovered by Seiji Ogawa in 1990. BOLD signals represent the energy emitted by hydrogen atoms in the brain when excited by a magnetic field. The strength of the signal is influenced by blood flow, blood volume and blood oxygenation. When the brain is activated, an increase in regional oxygenated

blood flow exceeds regional oxygen consumption, leading to an increase in signal intensity in regions activated by the task.

Task-based fMRI and resting-state fMRI

There are two types of data of fMRI: task-based fMRI and resting-state fMRI data. Task-based fMRI is used to detect activated areas of the brain. There are two common designs: block and event related. In a block design, external stimuli are presented in blocks and separated by short blocks of rest, while in an event-related design, stimuli occur spontaneously and have randomized time between events. Resting-state fMRI is used to measure regional connectivity when the brain is in resting state, without performing any explicit task [5].

Hemodynamic response functions (HRF)

BOLD signals are functional data whose relationship with stimuli can be characterized by the hemodynamic response functions (HRF). The HRF typically demonstrates the following patterns: the signal peaks approximately five seconds after stimulation and is followed by an undershoot that lasts a few seconds more. At high magnetic fields, an initial undershoot can sometimes be observed (Figure 2.1).

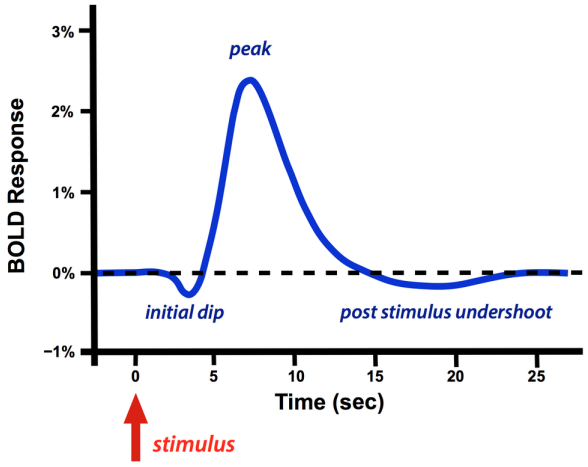


Figure 2.1: An Example of the Hemodynamic Response Function [10]

Voxel

An observed fMRI dataset contains a 3D array of intensity for each time point. A voxel is defined as a unit element of the 3D images. The data can be organized differently depending on the analytic tools and research questions. Here, we are more interested in the time courses of the voxels, since the activation of voxels are different for subjects with brain disorders including schizophrenia.

Preprocessing of fMRI data

A fMRI dataset is usually preprocessed in the following steps: skull stripping, correction for head motion, spatial smoothing, and registration to a template. After preprocessing, voxels are assumed to be functionally the same across subjects when they have the same spatial location. Given this assumption, we can save the 3D fMRI dataset in a matrix format, where each row contains the BOLD signal from a voxel, whose spatial location information is kept elsewhere (e.g. another matrix). Such data structure is more convenient especially when we are only interested in a small area of the brain.

Regions of Interest (ROI)

ROIs are specified regions in the brain of interest that are usually related to functional features or anatomic structures. While it is possible to study brain activation across all voxels, using only voxels in ROIs that are related to the stimuli requires less computation power and reduces the severity of multiple testing errors during statistical analysis [26]. For example, ROIs like striatum, thalamus, cuneus, precuneus, lateral and medial frontal eye fields (FEF), supplementary eye fields (SEF), and prefrontal cortex (PFC) have been shown to be consistently associated with anti-saccade tasks [9].

Challenge in fMRI data analysis

Given the above fMRI data's characterization, we are faced with the following difficulties in analyzing such type of data: (1) low signal to noise ratio (SNR), usually in the range of 2-4% [20] (2) moderate spatial and temporal correlation (3) computation difficulty due from having hundreds of voxels and hundreds of scans.

A desired algorithm should be able to address those problems.

2.1.1 Variability in the preprocessed fMRI data

In most previous research, variation of fMRI data is thought to stem from three aspects: drifts, noise and BOLD signals. Drifts are periodic noise from the patient's heartbeat or other systematic influence, while noise reflects intrinsic temporally and spatially correlated random variation. Variation in BOLD signals captures the variation of activation, a key aspect for detecting abnormality from brain disorders, and distinguishing diseased subjects given known brain areas which are influenced by the disorder.

However, new studies in brain connectivity for resting-state fMRI [14, 21] have shown that time lag exists across voxels and across subjects. It has been shown that accounting for temporal delays when inferring brain connectivity has substantially increased the node to node correlation and coherence.

While it is established that phase variability is a significant source of variation in fMRI data , few task-based fMRI studies have considered the subject-wise phase-variation of BOLD signals. It is only natural that the shape of the HRF differs by subject: some individuals may have a delayed peak compared to others, and some may have a more stretched out tail after the peak.

2.1.2 Motivating example

The study that motivates our method evaluated the effect of context on behavior and brain activity during saccade tasks among health subjects with high cognitive control, and low cognitive control and schizophrenia. FMRI data were collected while participants completed two runs in a block design: fixation alternating with anti-saccades (looking away from a cue). Specifically, a gray dot is presented on the screen, and subjects were asked to look at the dot when it was in the center and to look in the opposite direction when it appeared at a peripheral location.

The first run is referred to as "pre-practice" and the second run is referred to as "post-practice". Twenty ROIs are preselected based on the previous study of ROIs activated during saccade tasks [9]. Standard preprocessing steps were applied on the data including despiking, slice timing correction, masking (aligning functional images to anatomical images), warping

functional images to a standardized space, spatial smoothing and scaling to get rid of noised introduced by machine, anatomy differences, and head motion.

2.2 Elastic functional data analysis

To address the functional nature of the fMRI data and attempt to solve the problem of non-zero time lag in hemodynamic impulse response functions, we introduce the idea of elastic functional data analysis, a branch of FDA that focuses on the characterization of shapes of functions [33]. Elastic methods are often used in applications where arbitrary reparameterizations of the functions or curves of interest need to be considered, for instance the functions or curves of interest may not be in alignment. Elastic data analysis provides a different mean representation of curves than simple point-wise average and a special decomposition of variation in observed functions. Apart from classification, this framework can be used in a variety of statistical models including regression, principal component analysis, and clustering analysis. A comprehensive review of shape analysis/elastic data analysis can be found in [33].

2.2.1 Review of $\mathcal{L}^p$ space

Here we shortly review some properties of the the function space in the classic FDA framework and show why it is not an ideal framework to deal with functional data with temporal variability.

Let f to be a real-valued function defined on Ω, where Ω is a measurable space. The $\mathcal{L}^p$ space is a collection of functions that are up-to p degrees integrable, i.e. $\int_\Omega |f|^p d\mu < \infty$.

In FDA, it is commonly assumed that the functions are from the $\mathcal{L}^2$ space, where the $\mathcal{L}^2$ norm is defined as

$$\|f\|_2 = \left\{ \int_\Omega |f|^2 d\mu \right\}^{\frac{1}{2}}. \tag{2.1}$$

And the corresponding inner product of $f, g \in \mathcal{L}^2$ is

$$\langle f, g \rangle = \int_\Omega f \cdot g d\mu. \tag{2.2}$$

It can be shown that the $\mathcal{L}^2$ space defined above is a Hilbert space (i.e. a complete space with a metric induced by the norm), with the metric of form:

$$d(f, g) = \|f - g\|_2 = \sqrt{\int_0^1 |f(t) - g(t)|^2 dt}, \quad f, g \in \mathcal{L}^2. \tag{2.3}$$

For two functions in $\mathcal{L}^2([0, 1])$, warping is generally asymmetric: $d(f, g) \neq d(f \circ \gamma, g \circ \gamma)$.

Proof.

$$\begin{aligned} d(f,g)^2 &= \int_0^1 \left(f(s) - g(s)\right)^2 ds \\ &= \int_0^1 \left(f(\gamma(t)) - g(\gamma(t))\right)^2 \gamma'(t) dt && \text{(substituting variables)} \\ &\neq \int_0^1 \left\{f(\gamma(t) - g(\gamma(t))\right\}^2 dt && \text{(for most } \gamma(t)) \end{aligned}$$

□

2.2.2 Function representation in the Riemannian structure

In FDA, the objects of interest are functions which are often not fully observed. Usually, functional data (or "curve data") are considered to be a random sample of independent real-valued functions on a compact interval $I \subset \mathbb{R}$ with observation error, with functions observed at finitely many values. Without loss of generality, we restrict the interval to $I = [0, 1]$. It is convenient to assume that the observed functions (denoted as $f_i, i = 1, \cdots, N$) belong to a function space $\mathcal{F} = \{f : [0, 1] \to \mathbb{R} \mid f \text{ is absolutely continuous}\}$. We denote the vectorized observation of the functions as $f_i([t]) = (f(t_1), \cdots, f(t_T))$.

SRVFs: representation

For a given observed function $f \in \mathcal{F}$, we define its *square-root velocity function* (SRVF) representation as the transformation:

$$Q(f) : [0, 1] \to \mathbb{R}, \quad Q(f)(t) = \text{sign}(f'(t))\sqrt{|f'(t)|} \tag{2.4}$$

where f' denotes the derivative of f, and sign$(\cdot)$ is a sign function that outputs -1 when the input is negative, and 1 when the input is positive . For simplification, the SRVF representation is represented by q. It can be shown that $q \in \mathcal{L}^2([0, 1])$ and the original function f can be recovered by

$$Q^{-1}(q)(t) = f(0) + \int_0^t q(s)|q(s)|\, ds$$

which means the transformation is bijective up to a vertical translation. Since BOLD signals start at 0, i.e. $f(0) = 0$, the SRVF representation is unique for an observed fMRI function f.

The transformation space: a space of all phase variations

Depending on application context, there could be different classes of phase variations (i.e., shifting and scaling of the function f). Here, we assume that the space of warping of functions are the set of diffeomorphisms that are smooth and strictly increasing:

$$\Gamma = \{\gamma : [0,1] \to [0,1] \mid \gamma(0) = 0, \gamma(1) = 1, \dot{\gamma}(t) > 0\}. \tag{2.5}$$

Such group of time-warping functions restricts boundary behaviors and makes it possible to back-transform f for registration. Monotonicity is required because warping should not change the order of events, and the amount of smoothness is required to avoid discontinuous jumps in the time-warping functions, a practical assumptions for most applications.

This transformation space is a Lie group [33], where the identity element is $\gamma_{\text{id}}(t) = t$, and the inverse element γ^{-1} is the usual function inverse. In other literature, this space is sometime referred to as the *warping group*.

Similar to the observation function f, γ also has a unique SRVF representation: $\psi = Q(\gamma) = \sqrt{\gamma'}$. Note that the resulting SRVF space $\Psi^+ = \{Q(\gamma) \mid \gamma \in \Gamma\}$ is the positive orthant of the unit sphere in $\mathcal{L}^2([0,1])$, because ψ is positive and has $\mathcal{L}^2$ norm equal to 1. Here, $\mathcal{L}^2$ norm of the function has the form: $\|f\|_{\mathcal{L}^2} := \left(\int_0^1 |f|^2 \, \mathrm{d}t\right)^{1/2}$. We introduce the SRVF representation of the warping functions here but will defer further discussion on the

topic until next chapter, as the SRVF space Ψ^+ give us a nice description on the tangent space, which gets used with the Bayesian registration model.

The amplitude space

When using $\mathcal{L}^2$ norm to register functions in $\mathbb{F}$, one can find $\gamma_1, \gamma_2 \in \Gamma$ such that $f_1 \circ \gamma_1$ and $f_2 \circ \gamma_2$ are arbitrarily close, a phenomenon referred to as the "pinching effect". That is, one can find warping functions γ_1, γ_2 which arbitrarily "shrink" the time interval over which the two functions f_1 and f_2 are different, so that the distance between the warped functions are minimized. Here we refer to the example mentioned by Srivastava et al. (2016): in Figure 2.2, f_2 has two peaks, one of which is identical to f_1, a one-peak function, over the domain $[0, 0.6]$. No appropriate warping functions can match the two functions over $[0.6, 1]$, however, γ^* manages to pinch that interval of f_2, making the two functions identical.

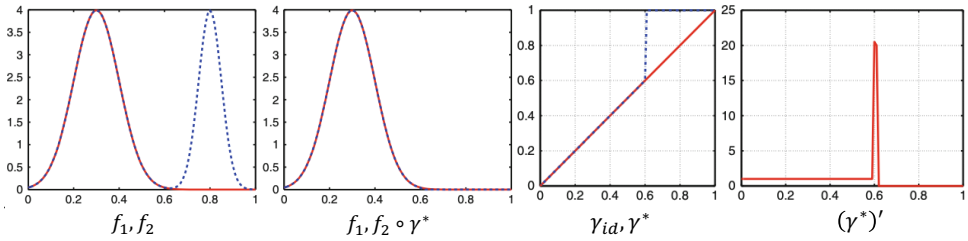


Figure 2.2: Illustration of pinching effect using $\mathcal{L}^2$ distance [33]

Furthermore, the $\mathcal{L}^2$ distance between two functions is not preserved after the transformation from the warping group, i.e. $\|f_1 - f_2\|_{L^2} \neq \|f_1 \circ \gamma - f_2 \circ \gamma\|_{L^2}$, so $\inf_{\gamma_1,\gamma_2\in\Gamma}\|f_1 \circ \gamma_1 - f_2 \circ \gamma_2\| \neq \inf_{\gamma_1\in\Gamma}\|f_1 \circ \gamma_1 - f_2\| \neq \inf_{\gamma_2\in\Gamma}\|f_1 - f_2 \circ \gamma_2\|$, resulting in the issue of asymmetry of registration. However, this is not the case for the SRVF representation.

When warped by γ, the SRVF representation of the warped function f becomes $Q(f \circ \gamma)(t) = Q(f)(\gamma(t))\sqrt{|\gamma'(t)|} := (q, \gamma)$. Then, it is clear that $\|q_1 - q_2\|_{\mathcal{L}_2} = \|(q_1, \gamma) - (q_2, \gamma)\|_{\mathcal{L}_2}$.

This property is the main reason that we use the SRVF representation. Given this property, the equivalence class of q induced by the warping group Γ, i.e. $[q] = \{(q, \gamma) \mid \gamma \in \Gamma\}$, represents a collection of functions in $\mathbb{F}$ that have the same "amplitude". Hence, the collection of all equivalence classes $[q]$, denoted by the quotient space $\mathcal{L}^2([0,1])/\Gamma$, is referred to as *amplitude space*.

2.2.3 Function alignment and registration

Here we introduce the elastic warping of FDA based on the framework laid out above. Specifically, it seeks to minimize the $\mathcal{L}^2$ distance of the SRVF representation. In the pairwise case, the formal registration problem is given by the following optimization:

$$\inf_{\gamma \in \Gamma} \|Q(f_1) - Q(f_2 \circ \gamma)\| = \inf_{\gamma \in \Gamma} \|Q(f_1 \circ \gamma) - Q(f_2)\|. \tag{2.6}$$

The infimum value is actually the distance between the two equivalence classes $[q_1]$ and $[q_2]$, and is a proper distance measure (i.e. it satisfies the properties of symmetry, positive definite, and triangle inequality). It can be shown that the inverse of the optimal γ on the left side is the optimal γ on the right side, a property called "inverse consistency" in area of image processing. In practice the optimal $\hat{\gamma}$ can be found using a dynamic programming algorithm (DPA), or a gradient-descent algorithm. One distinct difference between the two algorithms is that DPA finds the global optimum, while gradient-descent only finds local optima with no guarantees about finding the global optimum.

In the case of multiple registration, i.e. $N > 2$, the goal is to register all observed functions $f_1, f_2, \cdots, f_N$ to a common template via their corresponding equivalence classes

so that we can compare function amplitudes across population. This requires us to use the notion of Karcher mean of the warping functions ($\bar{\gamma}_N$):

$$\bar{\gamma}_N = \underset{\gamma \in \Gamma}{\operatorname{argmin}} \sum_{i=1}^{N} d_{\mathrm{FR}} \left(\gamma, \gamma_i\right)^2, \tag{2.7}$$

where d_{FR} is the Fisher-Rao metric of the transformation space Γ, which can be computed via

$$d_{\mathrm{FR}} \left(\gamma_1, \gamma_2\right) = d \left(\psi_1, \psi_2\right) = \cos^{-1} \left(\langle \psi_1, \psi_2 \rangle\right) = \cos^{-1} \left(\int_0^1 \psi_1(t) \psi_2(t) dt \right);$$

and the Karcher mean of amplitudes ($[\mu_q]$):

$$[\mu_q] = \underset{[q] \in \mathcal{A}}{\operatorname{arginf}} \sum_{i=1}^{N} d_a \left([q], [q_i]\right)^2 \tag{2.8}$$

where $\mathcal{A} = \mathcal{L}^2/\Gamma$ is the amplitude space.

Since $[\mu_q]$ represents a collection of functions in $\mathcal{F}$, it is sufficient to find a representative function of the equivalence class. One way to do this is to find the "center" of $[\mu_q]$, which makes the Karcher mean of all phases $\gamma_1, \cdots, \gamma_N$ which register observations to this template, equal to $\gamma_{id}(t) = t$.

The algorithm to find this center (also called "mean template")of $[\mu_q]$ first iterates between registering sample functions to the candidate $\tilde{\mu_q}$ and updating the candidate by taking the average of the registered SRVFs, then obtains the center of $[\tilde{\mu_q}]$ via warping $\tilde{\mu_q}$ by the inverse of $\bar{\gamma}_N$. An illustration of this process is shown in Figure 2.3:

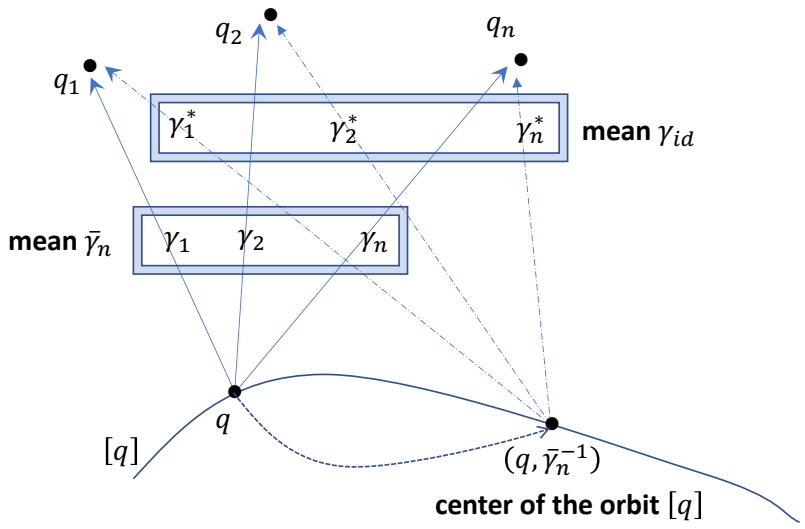


Figure 2.3: Illustration of finding the mean template in $[\mu_q]$ with respect to $(\gamma_1, \cdots, \gamma_n)$

Remark on fMRI signal registration:

We applied the multiple registration algorithm on the study of brain activity during blocks of antisaccade tasks among subjects with high cognitive control, low cognitive control and schizophrenia. After registration, the mean BOLD signals is able to capture the blocks of the external stimulus. Meanwhile, the amplitude variability in the samples is preserved in the warped mean BOLD signals while the cross-sectional mean has substantially lower level of peaks, a pitfall from averaging data outside of their normal modalities of variation.

2.3 Support vector machine (SVM)

SVM [15] is a classification method that produces significant accuracy with less computational power. The objective of the support vector machine algorithm is to find a hyperplane in the d-dimensional feature space that distinctly classifies the data points into two classes.

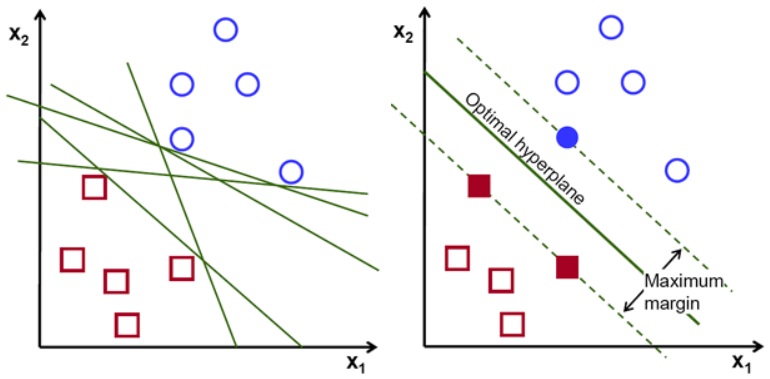


Figure 2.4: Illustration of the SVM algorithm

The optimal hyperplane found by SVM has the largest margin, the distance between data points (referred to as "supporting vectors") that are the closest to the hyperplane on each side. See Figure 2.4 for an illustration for a 2-dimensional classification problem. When data are not linearly separable, there are usually two types of modification that can be made. First, one can introduce regularization via slack variables so that we can have a soft margin that allows misclassification on a limited number of samples. Second, one can use kernel tricks that project the data feature space on a higher dimension. Popular kernels include Gaussian kernel and polynomial kernels.

Let y_i $(i = 1, \cdots, N)$ be the class of an observation $(y_i \in \{1, -1\})$, $\mathbf{x}_i$ be the d-dimensional feature vector, $\mathbf{w}$ be the coefficient vector of hyperplane and b be the intercept of the hyperplane. Then the optimization problem of the regularized SVM problem can be expressed

as:

$$\min \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \tag{2.9}$$

$$\text{s.t. } y_i\left(\mathbf{x}_i^T\mathbf{w} + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \forall i = 1, \cdots, N.$$

Here, ξ_i is the slack variable and $C > 0$ is the penalization parameter. This is called the regularized SVM problem, because it is equivalent to minimizing the hinge loss with a ridge penalty:

$$\min_{\mathbf{w},b} \frac{1}{N}\sum_{i=1}^{N}\max\left(0, 1 - y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right)\right) + \lambda\mathbf{w}^T\mathbf{w} \tag{2.10}$$

with $\lambda = \frac{1}{CN}$.

Suppose we can find a N-dimensional vector $\alpha = (\alpha_1, \cdots, \alpha_n)$ that satisfies the following Karush–Kuhn–Tucker (KKT) conditions:

$$\mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i,$$

$$0 = \sum_{i=1}^{N}\alpha_i y_i,$$

$$\alpha_i = C - \mu_i, \forall i$$

$$\alpha_i \left[y_i \left(\mathbf{x}_i^T \mathbf{w} + b \right) - (1 - \xi_i) \right] = 0, \forall i$$

$$\mu_i \xi_i = 0, \forall i$$

$$y_i \left(\mathbf{x}_i^T \mathbf{w} + b \right) - (1 - \xi_i) \geq 0, \forall i.$$

Then under KKT conditions, the optimization problem (2.9) (also called *primal problem*) is equivalent to the following problem, referred as the dual form of the original problem:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{N} \alpha_i + \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ &\text{s.t. } 0 \leq \alpha_i \leq C, \forall i = 1, \cdots, N \\ &\sum_{i=1}^{N} \alpha_i y_i = 0, \end{aligned} \tag{2.11}$$

where $\langle \cdot, \cdot \rangle$ is inner product in $\mathbb{R}^d$. Then, the classification function at x can be written as $\hat{C}(\mathbf{x}) = \sum_{i=1}^{N} \hat{\alpha}_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + \hat{b}$, where $\hat{b}$ can be solved by the KKT conditions. Finally, the data point $\mathbf{x}$ is classified as -1 if $\hat{C}(\mathbf{x}) < 0$, and 1 if $\hat{C}(\mathbf{x}) > 0$.

For nonlinear SVM problem, a transformation $h(\cdot)$ is applied to $\mathbf{x}$, which means $\mathbf{x}_i$ in the optimization problem (2.9) and (2.11) are then replaced with $h(\mathbf{x}_i)$. Essentially, this replaces the inner product in problem (2.11) by $K(\mathbf{x_i}, \mathbf{x_i}) = \langle h(\mathbf{x}_i), h(\mathbf{x}_j) \rangle$, where the inner product in the transformed space is properly defined (i.e. the transform space is a Hilbert space). Notice that this means it is not necessary to specify the transformation function $h(\cdot)$, as long as the kernel function $K(\cdot, \cdot)$ can be specified. A proper kernel corresponds to the inner product in the transformed space. By the Moore–Aronszajn theorem [1], $K(\cdot, \cdot)$ needs to be symmetric and positive definite.

2.3.1 Functional SVM (fSVM)

When the data are discretized observations of the underlying continuous functions in an infinite-dimensional space, functional SVM algorithms that take into account the functional nature of the data objects could substantially improve the performance of classification [30].

Let $f_i \in \mathcal{F}$ be the underlying function of the observed vector x_i, and t be a vector of the observed time points, then we can rewrite x_i as $f_i([t])$. The goal is to apply SVM on the unobserved functions.

It is possible to use SVM directly without applying any transformation/ kernel tricks, by approximating the inner product $\langle f_i, f_j \rangle_{\mathcal{F}}$ (e.g. $= \int f_j f_j \, d\mu$ for $\mathcal{F} = \mathcal{L}^2([0,1])$) by discretized observations. However, this brings out the issue of perfect fit, when the subspace of $\mathcal{F}$ spanned by the N observations is N-dimensional, which is a common case in practice. Hence, it is of interest to use adapted kernels for functional data in the framework of nonlinear SVM.

Functional kernels

It is possible to adapt the classical kernels to functions by replacing the inner product in $\mathbb{R}^d$ to the inner product in $\mathcal{F}$. However, that does not take into account the functional nature of the data.

A functional kernel that takes advantage of the functional nature can be defined via finite-dimensional projections of the functional space. Suppose the functional space is spanned by the basis system $\{\phi_1, \phi_2, \cdots\}$, then functions in the L-dimensional projection space ($L < N$) have the form of $\sum_{j=1}^{L} b_{ij}\phi_j$ where b_{ij}'s are the basis coefficients. It is easy to see that the inner product of this space $\left\langle \sum_{j=1}^{d} b_{ij}\phi_j, \sum_{j=1}^{L} b_{i'j}\phi_j \right\rangle = b_i^T \Phi b_{i'}$, where $b_i = (b_{i1}, b_{i2}, \cdots, b_{iL})$

and $\Phi = \left(\int \phi_i \phi_j d\mu\right)_{(i,j)}$. This gives us an adapted functional kernel:

$$K(f_i, f_j) = b_i^T \Phi b_j,$$

which we replace the inner product in (2.11) with, and then we have the following formal form of fSVM problem:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{N} \alpha_i + \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j b_i^T \Phi b_j \\ &\text{s.t. } 0 \leq \alpha_i \leq C, \forall i = 1, \cdots, N \\ &\sum_{i=1}^{N} \alpha_i y_i = 0. \end{aligned} \tag{2.12}$$

The classifier produced by fSVM has the form $\hat{C}(f) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(f_i, f) + \hat{b}$. Popular choice of basis system includes Fourier basis, wavelet basis, and splines.

In practice, the L-dimensional presentation of the observed discretized data is found via estimating of the basis coefficients by minimizing sum of squared errors on the observed time points.

2.4 Proposed approach for fMRI classification

To account for the spatial correlation in fMRI data, spatial fSVM has been developed by Samaddar (2019) [31]. Here we incorporate the ideas introduced by Samaddar to the elastic functional data analysis framework.

Our work uses the weighted structure of the kernel function proposed by Samaddar. Then we incorporate the Fisher-Rao metric as well as our model on how variability in the fMRI data should be decomposed into the construction of the kernel function. Finally, we propose the SVM algorithm in the elastic data analysis framework.

2.4.1 Spatial fSVM with function registration

Let $(\mathbb{X}_1, \cdots, \mathbb{X}_N)$ be a 3D fMRI dataset in matrix format, where each row contains the BOLD signal from one voxel, accompanied by a $S \times 3$ matrix $\mathbb{S} = (s_1, \cdots, s_S)^T$ that stores the coordinates of the S voxels in $\mathbb{X}_i = (x_{i1}, \cdots, x_{iS})^T$, and $(y_i, \cdots, y_N)$ where $y_i \in \{-1, 1\}$ be the corresponding classes of the N signal matrices.

Since voxels in different ROIs usually reflect different brain functions and consequently react to external stimuli differently, it is reasonable to assume that each ROI has its own template mean BOLD signal which is different from mean BOLD signal of another ROI in both amplitude and phase. In addition, we assume that BOLD signals from voxels in one ROI have additional variability in phase on the shared mean BOLD signal.

That is to say, for the observed data of voxel s in ROI r from subject i, denoted as $x_{i,s} = f_{i,s}([t])$, the underlying function $f_{i,s}(t)$ can be decomposed as:

$$f_{i,s}(t) = f_{roi=r}(t) \circ \gamma_{i,s}(t) + \epsilon_{i,s}(t) \tag{2.13}$$

where $\epsilon_{i,s}(t)$ is the combination of drifts and spatially correlated noises.

Then, it is assumed that the between group variation in BOLD signals is contained in the variation of the amplitude representation – the equivalence class $[q_{roi=r}]$ of $f_{roi=r}(t)$. Since

we are not concerned with the variability between different ROIs, and we do not intend to compare amplitude across ROIs, we will not continue to decompose the ROI-specific amplitude representations into the between-group variation and between-ROI variation.

In [31], the proposed spatial functional kernel of SVM is:

$$K\left(\mathbb{X}_i, \mathbb{X}_j\right) = \sum_s \sum_u w_{s,u} b_{i,s}^\top \Phi b_{j,u} \tag{2.14}$$

where

$$w_{s,u} = \begin{cases} 1, u = s \\ \rho^{d(s,u)}, d(s,u) < D_{\max} \\ 0, d(s,u) \geq D_{\max} \end{cases},$$

$b_{i,s}$ is a vector representation of projection of $f_{i,s}(t)$ in the L dimensional subspace with the basis system specified in Section 2.3.1, and Φ is a matrix of inner products of the basis functions. Here, ρ controls the spatial dependency, $d(u,s)$ measures the spatial distance between voxel u and voxel s, and $D_{\max}$ defines the neighborhood. The value of $D_{\max}$ is usually smaller than the radius of the ROI, such that $w_{s,u}$ is zero for voxels u and s that are not from the same ROI .

Essentially, (2.14) is a weighted sum of the functional kernels based on the relative distance between the voxels where the functions are from. The weight is designed to reflect the spatial correlation, in that the voxels that are closer to each other have a higher level of impact on each other's activation. Because the adapted kernel above is built upon the $\mathcal{L}^2$ distance of the observation space, it measures a mixture of phase and amplitude variability, when the only variability of interest is the amplitude variability.

Under the proposed model (2.13), we introduce a new adapted kernel induced by the elastic distance of functions, in the amplitude space for $f_{roi=r}$'s. A simple way of extracting the amplitude variability is to use the SRVF representation of the registered BOLD signals (registered to the ROI mean template) as the input data object for SVM. In this way, the transformed input will have less noise, and we would have a smaller error rate in classification.

Specifically, we:

1. find the L_f-dimensional representations $\hat{f}_{(i,s)}$ of the fMRI signals via a chosen basis expansion,

2. find the mean SRVF of each ROI given the L_f-dimensional representations: $\tilde{q}_r$,

3. register all L_f-dimensional representations of the fMRI signals to the corresponding ROI mean: $\tilde{x}_{i,s} = x_{i,s} \circ \gamma_{i,s}$, where $\hat{\gamma}_{i,s} = \text{arginf}_{\gamma \in \Gamma_I} \|\tilde{q}_r - (q_{i,s}, \gamma)\|$, and obtain the L_q-dimensional representation of the corresponding SRVF representation $\hat{q}_{i,s} = (q_{i,s}, \hat{\gamma}_{i,s})$

4. use cross-validation to find the optimal value of the hyper-parameters. Hyper-parameters include spatial dependency ρ, dimension L_q and L_f, regularization parameter $\lambda = \frac{1}{C}$, and the neighborhood distance $D_{\max}$.

Here the functional kernel is obtained via:

$$K(x_{i,s}, x_{j,u}) = \langle \hat{q}_{i,s}, \hat{q}_{j,u} \rangle. \tag{2.15}$$

where $\langle \cdot \rangle$ is the inner product of the L_q-dimensional space.

Suppose s and u are adjacent voxels in the same ROI while i and j are from different classes, then the BOLD signals are registered to the same mean template when the true

underlying amplitude should be different, i.e. $[q_{\{Y=1,roi=r\}}] \neq [q_{\{Y=-1,roi=r\}}]$. Hence, the kernel distance from (2.15) consists of the amplitude distance between the two classes.

One drawback of this proposed method is the high computation cost of registration: The dynamic programming algorithm (DPA) used in pairwise registration has the computation cost in the order of $O(kT^2)$ (T is the number of timepoints, $k \ll T$) according to [33], and each iteration step multiple function registration conducts DPA for all N samples, so the computation cost of registration alone grows drastically as the number of sample and number of time points increases. To speed up the registration process, we use subject-specific warping functions $\gamma_{i,ROI=r}$'s to replace the $\gamma_{i,s}$'s, so that functional signals from the same subject and same ROI share the same warping function. The subject-specific $\gamma_{i,ROI=r}$ of ROI=r is found by registering the point-wise average BOLD signals for each subject to a common mean.

Apart from being computationally efficient, using the simple point-wise average also mitigates the problem of over-fitting since averaging reduces the level of noise, which the registration algorithm is usually very sensitive to. Below we show an example of the noisy fMRI observations as well as two registration result: one based on subject point-wise mean of the fMRI data, one directly using the fMRI data. Here we present an example using a real fMRI study. (Data description is given in Section 2.5.2.) In Figure 2.5, the registered ROI mean function using the subject point-wise means has much clear patterns of the underlying block design of the study.

Then, given a new fMRI observation $\mathbb{X}_{\text{new}}$, subject-specific warping functions $\hat{\gamma}_{\text{new},r}$'s are found by registering point-wise mean BOLD signal in each ROI to the mean template SRVF $\tilde{q}_r$'s. Consequently, we have $K(\mathbb{X}_i, \mathbb{X}_{\text{new}}) = \sum_u \sum_s w_{s,u} \langle P_L(q_{i,s}, \hat{\gamma}_{i,s}), P_L(q_{\text{new},u}, \hat{\gamma}_{\text{new},u}) \rangle$, where

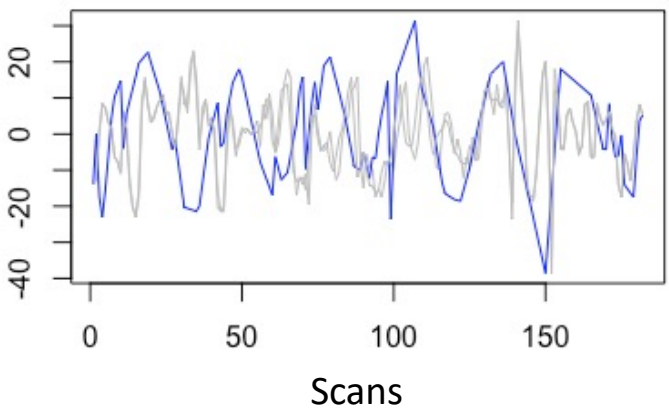


Figure 2.5: Comparison of registration by voxels. Point-wise means of the registered functions is presented for one subject in the ROI of rmedFEF. The blue line represents registration using subject point-wise means, and the grey line represents registration using individual voxels.

$\hat{\gamma}_{\text{new},u}$ is decided by the common warping function found for the ROI u belongs to. Finally, we can use the above kernel values to find the classifier $\hat{C}(\mathbb{X}_{\text{new}}) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(\mathbb{X}_i, \mathbb{X}_{\text{new}}) + \hat{b}$.

2.5 Simulation and real data applications

In the following sections, we compare the performance of the following four classification algorithms:

1. functional SVM using SRVF representation (fSVM),
2. functional SVM with registered data (rfSVM),
3. spatial functional SVM using SRVF representation (sfSVM),
4. spatial functional SVM with registered data (srfSVM).

2.5.1 Simulation study

Simulation setup

To mimic the BOLD signals, we use the signal and noise structure in the R package *neuRosim* [43]: we allow signals to follow different stimulus and noises of a mixture of white noise, colored noise, temporal noise and spatial noise. Note that this simulated data is "preprocessed" in the way that artifacts like head motion and inhomogeneity in the magnetic field are not explicitly modeled.

We use $N = 50$ for the number of subjects in each group throughout the simulation, and repeat each simulation setting 50 times. In general, the image space is 2-dimensional (50×50) for easier visualization, temporal solution (TR)=2 s, and the stimulus is an ON/OFF block design of 200s. Then, the activation in BOLD signals is modeled as the convolution of a HRF and the stimulus vector, and the shape of HRF is modeled via the double-gamma HRF, which models an initial dip and an undershoot after the peak in the BOLD signals.

Next, we present the framework of modeling the variability in BOLD signals.

(1) Between-ROI variability: ROIs are defined as activated areas with predefined centers and radius, and the shape of HRFs across ROIs varies by different parameter setup of the double-gamma HRF.

(2) Between-group variability: the between-group variability is a combination of the difference in effect sizes of the activation and in the shape of the double-gamma HRFs.

(3) Phase variability: activation across voxels and subjects is simulated by random warping of the mean ROI BOLD signal, where the random warping function can be generated by the R package *fdasrvf* [38].

(4) Noise structure: the noise is a mixture of white noise, colored noise, temporal noise and spatial noise. Specifically, we specify the signal-to-noise ratio, the correlation of the autoregressive temporal noise, and the FWHM (the full-width-half-maximum of the Gaussian kernel) of the Gaussian random field for spatial noise. We use the default setup for the rest of the noise components.

Simulated data I

Given the simulation structure, we specify a two-group fMRI data using setup specified in Table 2.1 below:

Table 2.1: Simulation setup I

	Group 1	Group 2
General setups		
effect size	1	1.2
signal-to-noise ratio (SNR)	2	
double gamma parameters	$a_1 = 6$, $a_2 = 12$, $b_1 = 0.9$, $c = 0.35$	
	$b_2 = 1.3$	$b_2 = 0.9$
ROI 1		
number of voxels	49	
onsets (s)	1, 50, 100, 150	
duration of onsets	20	
ROI 2		
number of voxels	49	
onsets (s)	15, 65, 115, 165	
duration of onsets	10	

The visualization of one simulated example is shown in Figure 2.6 and Figure 2.7.

Classification and results

To obtain better estimation of the warping functions, we allow the Fourier approximation of the simulated data to have period equal to the period in the block design. And then, registration is done only with the function within the period.

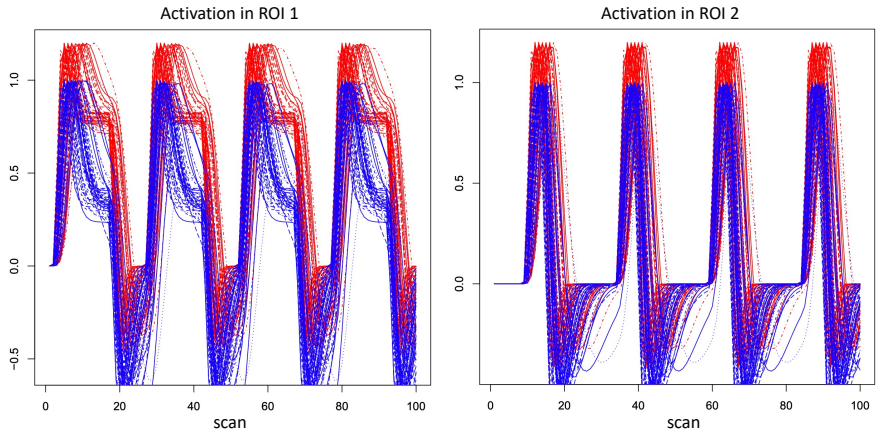


Figure 2.6: Activation signals from setup I

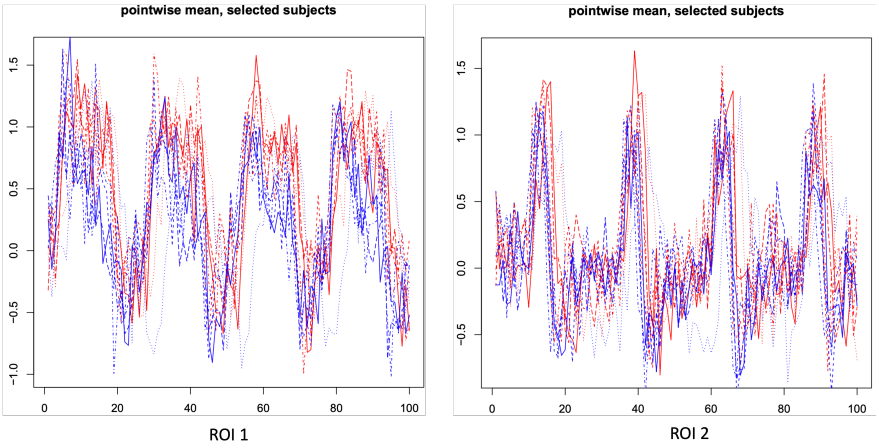


Figure 2.7: Point-wise average of the simulated fMRI data I

Finally, we tune the parameters mentioned in Section 2.4.1 via 5-fold cross-validation. We tune ρ in the range $(0, 1)$, λ in the set $\{1, 0.1, 0.01, 0.001, 0.0001\}$, L_f in $[12, 17]$, L_q in $[12, 17]$, and $D_{\max}$ among the 40^{th}, 60^{th}, and 80^{th} percentile of pairwise distances of voxels that belong to the same ROI. Specific values used in the cross-validation is shown in Table 2.2. The data are split into a training set (80%) and a testing set (20%) for performance evaluation.

Table 2.2: Hyper parameters for the simulation study

Hyper parameters	values considered
ρ	$0.1, 0.2, \cdots, 1$
λ	0.0001, 0.001, 0.01, 0.1, 1
L_f (pairs)	3 - 7
L_q (pairs)	5 - 11
$D_{\max}$	3 - 7
fourier period	4

The registration results are shown in Figure 2.8. Comparing the subject-and-ROI specific signals before and after registration, we can see that our registration step have successfully separated the random warping from the general variability in the functions.

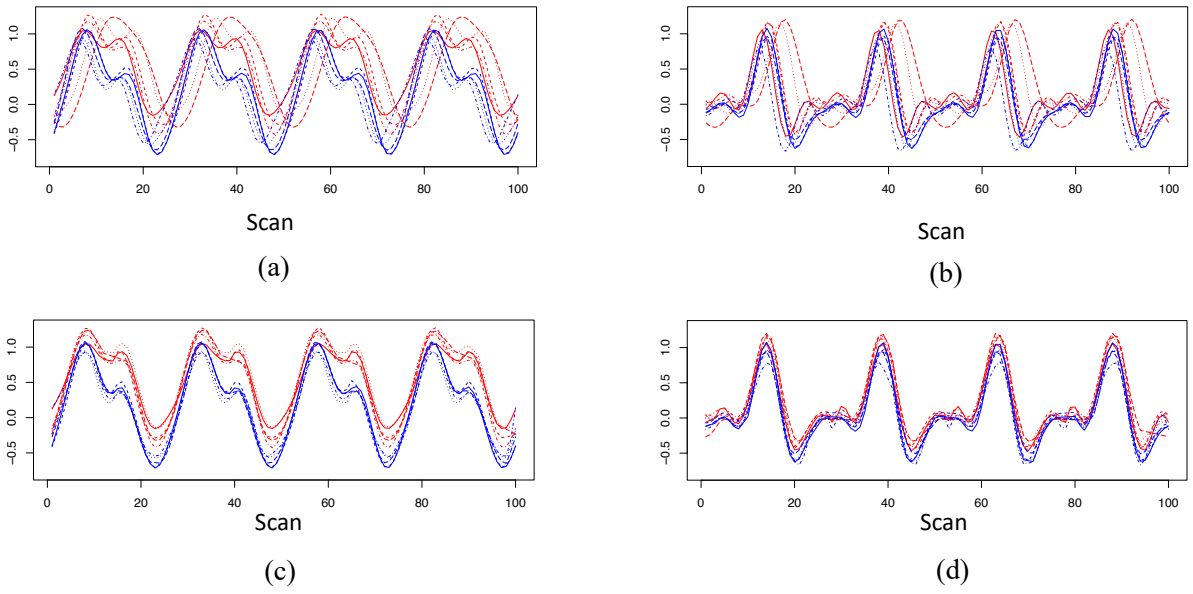


Figure 2.8: Registration of the functions in simulated data I. The functions are color-coded based on their group assignment: (a) and (b) are point-wise mean of the smoothed raw data for selected subjects from the two ROIs, and (c) and (d) are the point-wise mean of the registered version for those subjects.

The classification performances of the four algorithms on this simulation setup are shown in the boxplot below (Figure 2.9). From the classification results, we can see that both registration and using spatial information helps distinguish the two groups of subjects and the srfSVM algorithm has the lowest median test error rate, even though the spatial algorithms (sfSVM and srfSVM) shows sign of over-fitting.

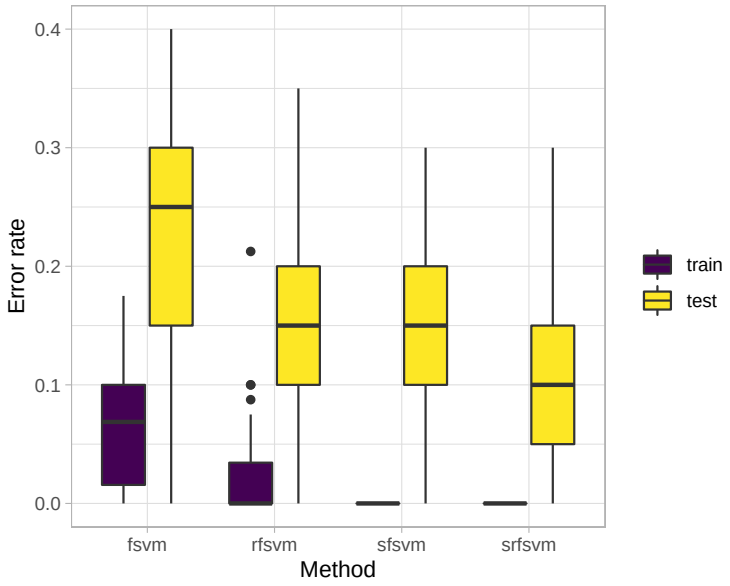


Figure 2.9: Performance comparison based on the simulated data I

Simulated data II

Another simulation is constructed based on the parameters specified in Table 2.3. This simulation setup is the same as the previous one except that the double gamma parameters are restricted to the same between the two groups and the signal-to-noise ratio (SNR) is increased to 5 from 2.

The visualization of one simulated example from the second simulation setup (Table 2.3) is shown in Figure 2.10 and Figure 2.11.

Table 2.3: Simulation setup II

	Group 1	Group 2
General setups		
effect size	1	1.2
signal-to-noise ratio (SNR)	5	
double gamma parameters	$a_1 = 6$, $a_2 = 12$, $b_1 = 0.9$, $c = 0.35$	
	$b_2 = 1.3$	$b_2 = 1.3$
ROI 1		
number of voxels	49	
onsets (s)	1, 50, 100, 150	
duration of onsets	20	
ROI 2		
number of voxels	49	
onsets (s)	15, 65, 115, 165	
duration of onsets	10	

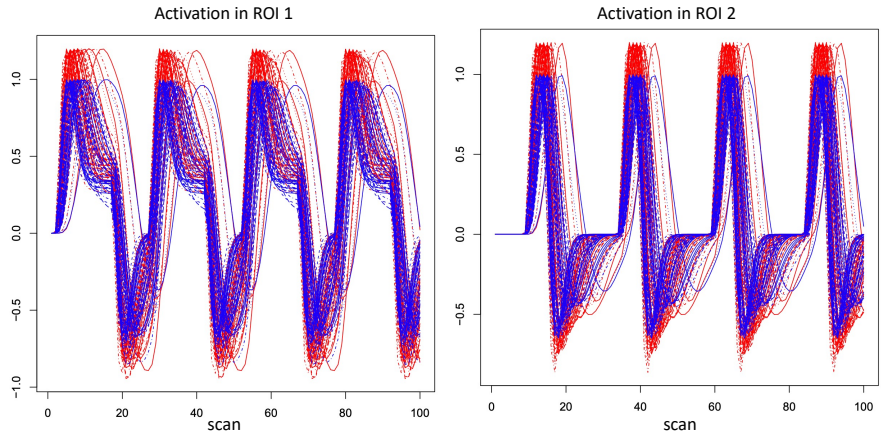


Figure 2.10: Activation signals

Classification and results

Because the spatial structure and design of the fMRI study are the same between the two simulated data, we adopt the same specifications in the cross-validation step for tuning those hyper-parameters. Again, the data are split into a training set (80%) and a testing set (20%) for performance evaluation.

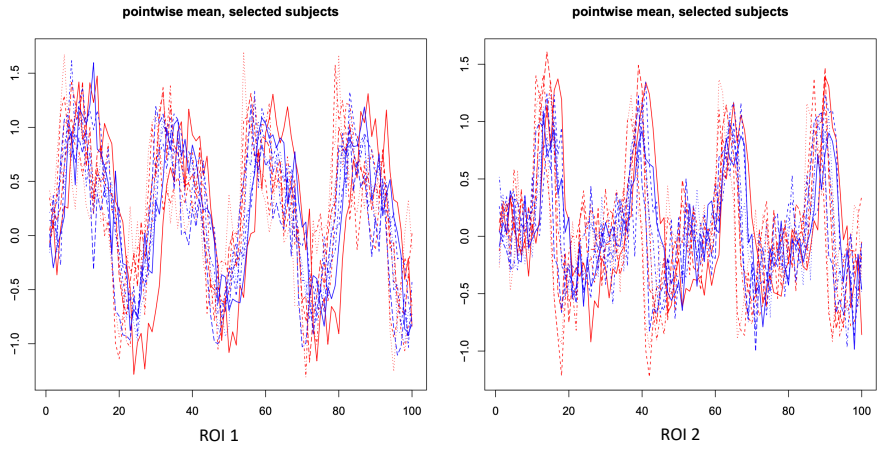


Figure 2.11: Point-wise average of the simulated fMRI data

The registration results are shown in Figure 2.12. Comparing the subject-and-ROI specific signals before and after registration, we still see a clear separation of the random warping from the general variability in the functions by taking a registration step. Because the HRF parameters are the same between the two groups, we expect that registration in this simulation setup will make even more difference in the classification.

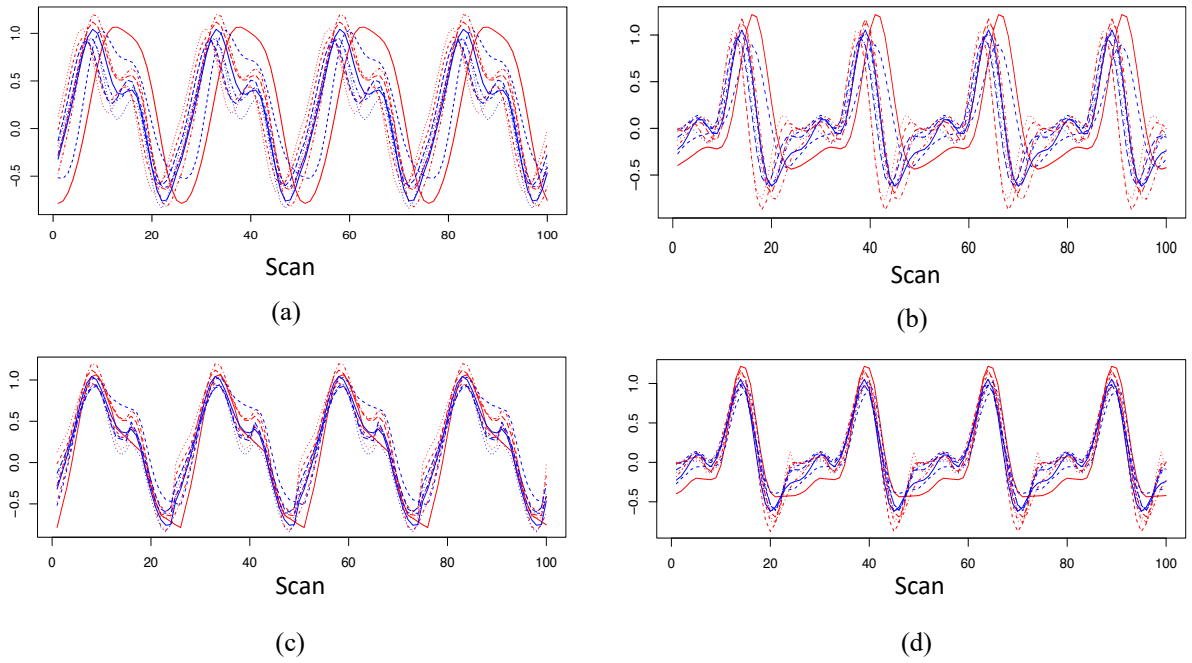


Figure 2.12: Registration of the functions in simulated data II. the functions are color coded based on their group assignment. (a) and (b) are point-wise mean of the smoothed raw data for selected subjects from the two ROIs. (c) and (d) are the point-wise mean of the registered version for those subjects.

The classification performances of the four algorithms on simulated data II are shown in the boxplot below (Figure 2.13). From the classification results, we also see that both registration and using spatial information helps distinguish the two groups of subjects. Apart from the discoveries we found in simulated data I, we also see that the SVM algorithms that involve a registration step have considerable better performance on the test set compared to the SVM algorithms without such a step. This is consistent with our expectation, since

now the shape of the double-gamma function is set to be the same between the two groups, leading to less group-wise variability that can be discovered just with the raw fMRI data.

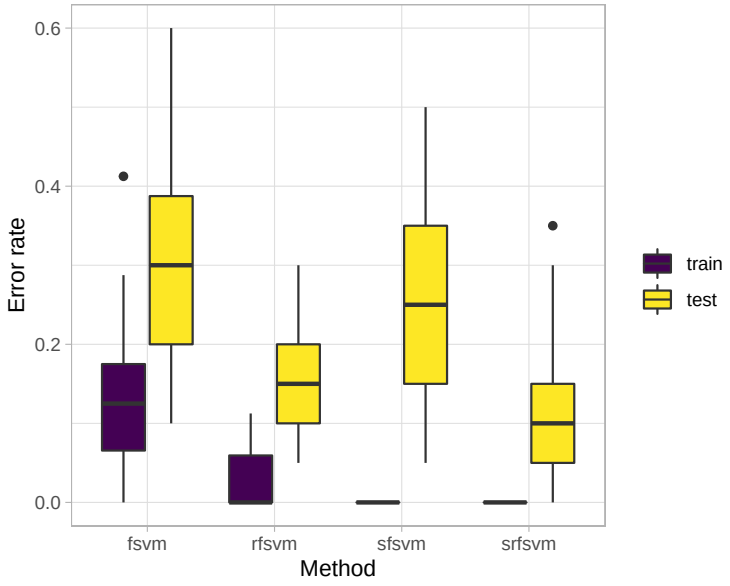


Figure 2.13: Performance comparison based on the simulated data II

2.5.2 Real data analysis

Data description

Healthy participants recruited from the community were categorized into two groups based on pretest cognitive control scores: subjects scoring in the top third of the distribution were assigned to the high cognitive control group ($N = 31$) and subjects scoring in the lower third of the distribution were assigned to the low cognitive control group ($N = 30$). FMRI data were then acquired while subjects were engaged in blocks of antisaccade task.

The study was conducted at the University of Georgia Bio-Imaging Research Center with a GE Excite HD 3.0 T MRI scanner. Two runs of an antisaccade task were performed during fMRI in an event-related design: antisaccade stimuli were presented in the center of the screen between 2,300 and 6,000 ms, disappeared for 200 ms, and reappeared at a peripheral location, at which time subjects had 1,000 ms to make a response. The corresponding fMRI data were acquired with the T2*-weight scan, which resulted in 33 slices of image matrix of 64×64, with FOV = 22 cm and slice thickness = 4 mm that obtained with TE = 30 ms, TR = 2000 ms. Standard preprocessing steps were taken following the same protocol used in the motivating study.

Finally, the functional signal observed at 182 timepoints from 1666 voxels are selected from the 20 ROIs (Table 2.4) selected by [9]. To assess classification performance, we randomly divide each group into training (N=41) and testing (N=20) sets for 10 runs.

Table 2.4: 20 pre-specified ROIs

Region	Abbreviation	Number of voxels in ROI
Supplementary Eye Field	SEF	80
Right Thalamus	rthal	81
Right Precuneus	rprecuneus	87
Right PreFrontal Cortex	rPFC	76
Right Middle Occipital Gyrus	rMOG	82
Right Medial Frontal Eye Field	rmedFEF	85
Right Lateral Frontal Eye Field	rlatFEF	83
Right Inferior Parietal Lobe	rIPL	84
Right Left Frontal Cortex	rlFC	87
Right Cuneus	rcuneus	80
Right Basal Ganglia	rBG	73
Left Thalamus	lthal	84
Left Precuneus	lprecuneus	92
Left PreFrontal Cortex	lPFC	80
Left Middle Occipital Gyrus	lMOG	91
Left Medial Frontal Eye Field	lmedFEF	85
Left Lateral Frontal Eye Field	llatFEF	89
Left Inferior Parietal Lobe	lIPL	76
Left Cuneus	lcuneus	80
Left Basal Ganglia	lBG	73

Classification and results

The design of this fMRI study is not given. We tune the parameters mentioned in Section 2.4.1 via 5-fold cross-validation. Ranges of those parameters are given in Table 2.5.

Table 2.5: Hyper parameters for the real data

Hyper parameters	values considered
ρ	$0.1, 0.2, \cdots, 1$
λ	0.0001,0.001, 0.01, 0.1, 1
$D_{\max}$	4-12
L_f (pairs)	6 - 10
L_q (pairs)	8 - 15
fourier period	1

The classification performance of the four algorithms on one of the 10 splits are shown in the table below (Table. **??**). The regular fSVM algorithm is no better than random guess

when it comes to the validation data, while all the other three algorithms has better performance. This real data application confirms our assumption that separating the amplitude and phase variability as well as incorporating spatial information help identify the functional characteristics regarding the underlying brain disorder.

Table 2.6: Algorithm performance on the real data

Algorithms	training error (N = 41)	test error (N=20)
fSVM	0.073	0.5
rfSVM	0.0487	0.40
sfSVM	0	0.45
srfSVM	0	0.35

Here we also show the mean classification errors from the 10 runs in Table 2.7.

Table 2.7: Mean classification errors on the real data among 10 runs

Algorithms	training error (N = 41)		test error (N = 41)	
	Mean	Std. Dev.	Mean	Std. Dev
fSVM	0.083	0.11	0.52	0.15
rfSVM	0.092	0.15	0.45	0.10
sfSVM	0	0	0.4	0.07
srfSVM	0	0	0.37	0.06

2.6 Discussion

On registration:

Even though the method we propose for fMRI data is an improvement of the existing method, it has the following flaws that could be improved:

1. The registration based on SRVFs involves first derivatives of functions, so there is a strong possibility of noise influencing the alignment results. Under low signal-to-noise ratio, the alignment is more governed by noise than signal, and in turn, registered functions may not capture the population-level feature of the amplitude space well, especially when the

added noised generate artificial "bumps" in the observed functional data, which makes the algorithm search warping functions that match up the noise "bumps" with the signal-related oscillations. We try to control this a little bit by taking point-wise averages within ROIs prior to using SRVF-based registration, which is an ad hoc solution that might not work well with other fMRI data. Srivastava et al. showed in [33] that pre-smoothing of the functional data via wavelet-based denoising improves the registration result significantly for simple Gaussian noise structure.

2. The kernel matrices of the adapted kernels are dense and full rank, which complicates the implementation of quadratic programming algorithms. The algorithm becomes time-consuming when the training data size is large, the number of voxels selected are large, and the number of basis increases. Some low-rank approximation techniques could be used to speed up the algorithm, or we simply narrow down on which voxels to select. In addition, we find that the proposed algorithm tends to overfit the training data, a problem which can only be mitigated via using very specific ranges of values for hyper-parameters during cross-validation.

3. An observation we found during simulation setup is that when the period of the activation signals increases, the classification error also increases. This is consistent with the findings on alignment of "quasi-periodic" time-series like ECG data in [6]. Due to the periodic design of fMRI studies, the registration results could be improved by the quasi-periodic dynamic time warping proposed in [6], especially when the design is event-related, i.e. the data are "quasi-periodic".

On classification:

Over the years, other classification methods (e.g. [40]) have been proposed specifically for elastic functional data analysis. Given their discussion, the following issues can be considered on the fMRI classification task:

1. A more complex model that incorporates amplitude variability, phase variability and variability in the initial point should be considered. The phase variability could rise from the nature of the fMRI signals, and the variability in initial point should be considered since SRVF representation of a function $f(t)$ is unique only up to the constant $f(0)$. In the simulation study, we specifically assumed that the underlying activation is constant at start, and the phase variability is purely random, so our proposed methods works. However, those assumptions are not necessarily true. There are function representations proposed recently that concatenates the three variability components and measure the overall variability based on a new Fisher-Rao metric on the new representation space.

2. A generative model which output probabilities instead of classification should be considered. In reality, brain disorders are rare events, so it could be more useful if doctors can be provided with a specific risk estimation instead classification predictions. There have been methods proposed (e.g. [39]) that estimate probabilities instead of classification result using the framework of logistic regression. This type of model framework can be applied to the spatially-dependent fMRI data.

Chapter 3

Bayesian framework on the spatial temporal data

Various models for the Bayesian registration of functional data have been developed recently. Like all Bayesian methods, Bayesian registration has the benefit of incorporating prior information, and probability-based conclusions via posterior distributions. However, most Bayesian registration proposed assume a priori independence of functions and their warping.

In the case of spatial temporal data like the fMRI data we investigated in the previous chapter, spatial information can be incorporated via prior distributions. In this chapter, we address the spatial dependence among functions and consider extending one Bayesian registration model proposed by Lu, Herbei and Kurtek [23] to account for spatial dependence in warping functions.

The rest of this chapter is organized as follows. Section 3.1 gives a brief introduction on the existing Bayesian registration method, Section 3.2 explains the Bayesian framework of

function registration proposed in [23], and Section 3.3 describes the proposed prior model used in the special case of spatially-indexed functional data as well as the MCMC algorithm used to obtain the posteriors. Finally, we show a simulation application in Section 3.4, a real data application in Section 3.5 and have a brief discussion on the proposed method in Section 3.6.

3.1 Review of Bayesian registration methods

One of the first Bayesian hierarchical registration models is proposed by Telesca et al. [36]. Their model was not based on the framework of Elastic Data Analysis, but rather based on the shape-invariant model (SIM) [13]. The underlying subject-specific amplitude functions are random linear transformation of a common smooth curve, and the subject-specific warping functions are modeled using a pre-defined basis system and a multivariate normal distribution of the basis coefficients after imposing the monotone and boundary conditions for the warping function. They specifically investigated the application using B-spline basis. For the posteriors, the samples are drawn based on a "Gibbs-within-Metropolis" algorithm, where the parameter is updated using a Metropolis-Hastings step and a uniform proposal instead of a Gibbs step, when analytical form of the full conditional distribution is unavailable. This "Gibbs-within-Metropolis" algorithm is a prevalent trick used in Bayesian registration problems.

In recent years, methods of Bayesian registration based on the framework of Elastic Data Analysis have been proposed. Working with the SRVF representations enables the method to be invariant under simultaneous warping of all the curves, but this is not the case in [36].

Most of the methods based on the SRVF representation utilize the Gaussian process model in the amplitude space, but they differ in the proposed models of the warping functions.

For example, Cheng et al.[7] suggested a Dirichlet process on the warping functions directly, since the warping functions are indeed cumulative distribution functions. Specifically, the gaps $\gamma(t_i) - \gamma(t_{i+1})$ follow a Dirichlet distribution. In this way, they avoided the need of a basis system, and were able to focus the spherical warping space directly.

However, using a Dirichlet process prior makes the model highly depend on the discretization of the warping functions, since the distribution is modeled based on the gaps of consecutive values of those warping functions. Lu et al.[23] proposed that by mapping the warping functions to their equivalence in the tangent space, they can avoid the complexity about choosing different discretization by proposing Gaussian process priors on the tangent space of the warping functions, which is a linear space, directly. Given their prior setup, prior knowledge can be incorporated to control the smoothness as well concentration level of the warping functions. Their method can be applied to data observed on a fine gridding since it is robust to different grid sizes of the functions.

Given the advantages of the method proposed by Lu et al., we adapt this Bayesian registration method for spatially dependent functional data.

3.2 Bayesian registration via Gaussian process priors

In this section, we briefly introduce the framework proposed by [23], using the same notation and setup discussed in Section 2.2.

The energy-based elastic function registration framework (Section 2.2) estimates warping functions $\gamma \in \Gamma$ via dynamic programming by solving the optimization problem in (2.6). In this Bayesian framework, estimation of the warping functions is achieved by computing (or sampling from) the posterior distribution given the data, a Gaussian process prior on the amplitude space, and a series of independent Gaussian process priors on a transformed space of the warping function space Γ.

This transformation of the space of γ exploits the Riemannian-geometric property of the warping function space defined in (2.5). As explained in Section 2.2, the SRVF space of warping functions $\Psi^+ = \{Q(\gamma) \mid \gamma \in \Gamma\}$ is a positive orthant of the unit sphere of $\mathcal{L}_2([0,1])$, where arclength distance is equivalent to the Fisher-Rao metric on the warping functions.

Although the SRVF representation Ψ^+ of the warping group simplifies the complicated geometry of Γ, it is still nonlinear. To deal with the nonliearity, the inverse of exponential map is taken so that the unit sphere Ψ is unwrapped on to the tangent space, denoted as $T_1(\Psi)$, at the constant function $1 \in \Psi$ (which represents the identity function $\gamma_{id} \in \Gamma$):

$$T_1(\Psi) = \left\{ g : [0,1] \mapsto \mathbb{R} \mid \langle g, 1 \rangle = \int_0^1 g(t)dt = 0 \right\}.$$

The exponential map and inverse exponential map can be explicitly expressed:

$$\begin{aligned} &\exp_1 : T_1(\Psi) \mapsto \Psi \quad \exp_1(g) = \cos(\|g\|) + \frac{\sin(\|g\|)}{\|g\|} g, \quad g \in T_1(\Psi), \\ &\exp_1^{-1} : \Psi \mapsto T_1(\Psi) \quad \exp_1^{-1}(\psi) = \frac{\theta}{\sin(\theta)}(\psi - \cos(\theta)), \quad \theta = d(1,\psi), \quad \psi \in \Psi, \end{aligned} \tag{3.1}$$

where $d(1,\psi) = \cos^{-1}\left(\int_0^1 \psi(t)dt\right)$ is the arclength distance on the unit sphere Ψ. The transformations mentioned above are illustrated in Figure 3.1.

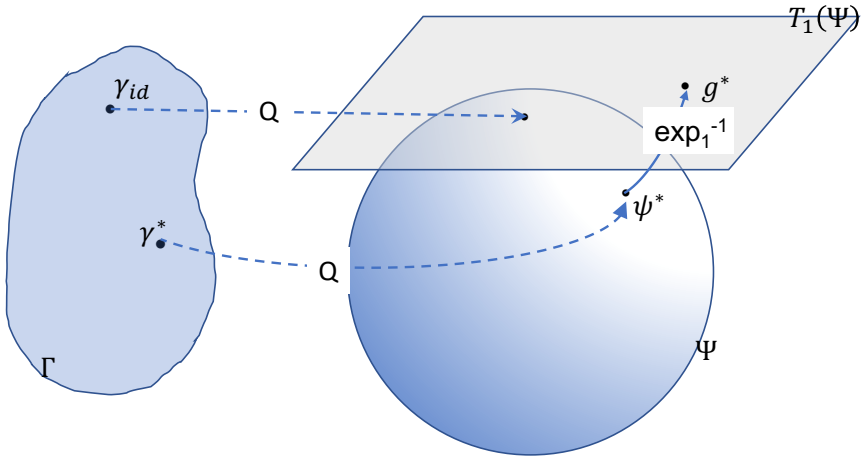


Figure 3.1: Illustration of the transformations of the warping space

The tangent space $T_1(\Psi)$ is a linear space and hence Gaussian process prior can be assigned on such space. Note that when mapping $T_1(\Psi)$ back to Ψ^+ via $\exp_1^{-1}$, an additional truncation step need to be taken. Consequently, the Gaussian prior is actually a truncated Gaussian process such that the domain mapped back to the sphere Ψ lies within the positive orthant Ψ^+. For observation representation, we again adopt the SRVF representations: $\mathcal{L}^2 = \{q = Q(f) | f \in \mathbb{F}\}]$.

3.2.1 Bayesian model specification

To adopt the idea of the Karcher mean of the amplitude space, the model for sampling distribution is built based on the discretized SRVF functions $q_i([t])$'s and a multivariate Gaussian distribution, so that Bayesian likelihood is a function of the $\mathcal{L}_2$ distance between SRVF of observation and that of the mean template.

In the case of multiple registration, we need to estimate both the subject-specific warping functions γ_i's and the template SRVF function q^* that all q_i's are registered to. Thus, we also assign a Gaussian process prior with this unknown template q^*.

The proposed Bayesian model can be specified as below. Let the vectorized observation of the functions be $f_i([t]) = (f(t_1), \cdots, f(t_T)), i = 1, \cdots, N$, and a subset $A \subset T_1(\Psi)$ be defined as:

$$g \in A \subset T_1(\Psi) \Longleftrightarrow \exp_1(g) > 0. \tag{3.2}$$

Model 1:

Prior:

$$\begin{aligned} g_1, \ldots, g_N &\sim \text{ Gaussian}\left(0, \mathcal{C}_g; I_A\right) \\ q^* &\sim \text{ Gaussian}\left(0, \mathcal{C}_q\right), \text{i.i.d} \\ \sigma^2 &\sim \text{IG}(\text{ shape } = a, \text{ scale } = b) \end{aligned} \tag{3.3}$$

where the joint prior is the independent joint distribution of the specified priors above.

Sampling distribution:

$$\{(q_i, \gamma_i)([t])\}_{i=1}^N \mid \{\gamma_i\}_{i=1}^N, \sigma^2, q^* \sim MVN(q^*([t]), \sigma^2 I_N), \text{i.i.d.} \tag{3.4}$$

where γ_i and g_i is linked by

$$\gamma_i = Q^{-1}(\exp_1^{-1}(g_i)). \tag{3.5}$$

Here, Gaussian $(\cdot, \cdot; I_A)$ refers to the Gaussian process restricted to the domain A, $IG(\text{shape}, \text{scale})$ refers to the inverse gamma distribution, and MVN refers to the multivariate normal distribution. Also, $\mathcal{C}_g, \mathcal{C}_q$ are known covariance operators, and a, b are known constants.

The posterior distribution of $\{g_1, \cdots, g_N; q^*; \sigma^2\}$ is obtained using a Metropolis-within-Gibbs (MwG) algorithm [8]. That is, at each step, we sample from the full conditional distribution of each parameter iteratively. The function parameters are updated via a Metropolis step and σ^2 is updated via a simple Gibbs step from the Inverse Gamma distribution (3.7). The likelihood of this model has the form:

$$L\left(\{g_i\}_{i=1}^N, \sigma^2, q^* \mid q_1, \cdots, q_N\right) \propto \prod_{i=1}^{N} \left(\frac{1}{\sigma^2}\right)^{T/2} \exp\left\{-\frac{1}{2\sigma^2} \operatorname{SSE}(g_i, q^*)\right\}, \tag{3.6}$$

and the full conditional distribution of σ^2 is

$$\sigma^2 \mid \{q_i\}_{i=1}^N, \{g_i\}_{i=1}^N, q^* \sim \text{IG}\left(\text{ shape } = \frac{1}{2}NT + a, \text{ scale } = \frac{1}{2}\sum_{i=1}^{N} \operatorname{SSE}\left(g_i, q^*\right) + b\right), \tag{3.7}$$

where

$$\operatorname{SSE}(g_i, q^*) = \sum_{j=1}^{T} \left(q^*\left(t_j\right) - q_i\left(\int_0^{t_j} \exp_1^2(g_i)(s)ds\right) \times \exp_1(g_i)\left(t_j\right)\right)^2. \tag{3.8}$$

The specific MCMC steps used for the Bayesian registration can be found in the complementary materials in [22]. Below, we only introduce the key components in the updating process.

3.2.2 Z-mixture Preconditioned Crank-Nicolson (pCN) proposal

Here we lay out the specific Metropolis-Hastings (M-H) algorithm used for updating the function parameters. M-H algorithm is a popular MCMC method, which has the form of proposing candidate samples and accepting the candidates with some probability. Specifics of M-H algorithms can be found in any Bayes textbook (e.g. [12]).

The key component of a M-H algorithm is the specific proposal used to generate candidate samples. Here, the Z-mixture pCN proposal is used to update the function parameters $\{g_i\}$ and q^*. This proposal enjoys the benefit of (1) being independent of the dimension of the discretized functional data, which is far more efficient when the data are observed under a fine grid compared to a simple random walk proposal; and (2) having simple form of the M-H acceptance ratio which does not require a closed form expression of the distribution of functions. The general form of the Z-mixture pCN proposal can be found in [22]. Here, we describe the specific application of this proposal in updating $\{g_i\}$ and q^*.

2-mixture pCN algorithm for updating $\{g_i\}$ and q^*

- Update $g_i^{(m)}$ to $g_i^{(m+1)}$ iteratively, given all the other parameters at the current state (m):

  1. Propose

$$g' = \begin{cases} g_i^{(m)}\sqrt{(1-\beta_1^2)} + \beta_1\xi & \text{with probability } p_1 \\ g_i^{(m)}\sqrt{(1-\beta_2^2)} + \beta_2\xi & \text{with probability } p_2 \end{cases} \tag{3.9}$$

  where $p_1 + p_2 = 1$, $\epsilon \sim \text{Gaussian}\,(0, \mathcal{C}_g)$. Re-propose if $\exp_1(g') \leq 0$.

2. Accept g' as $g_i^{(m+1)}$ with probability

$$\min\left\{1, \exp\left(\ell\left(g', \cdots, \left(\sigma^2\right)^{(m)}\right) - \ell\left(g_i^{(m)}, \cdots, \left(\sigma^2\right)^{(m)}\right)\right)\right\}, \tag{3.10}$$

where the likelihood ratio can be simplified to

$$\exp\left(-\frac{1}{2\sigma^{2(m)}}\left(\mathrm{SSE}\left(g_i', q^{*(m)}\right) - \mathrm{SSE}\left(g_i^{(m)}, q^{*(m)}\right)\right)\right), \tag{3.11}$$

where $\mathrm{SSE}(\cdot,\cdot)$ is the same as (3.8).

- Update $q^{*(m)}$ to $q^{*(m+1)}$ given all the other parameters at their state:

1. Propose

$$q' = \begin{cases} q^{*(m)}\sqrt{(1-\beta_1^2)} + \beta_1\xi & \text{with probability } p_1 \\ q^{*(m)}\sqrt{(1-\beta_2^2)} + \beta_2\xi & \text{with probability } p_2 \end{cases} \tag{3.12}$$

where $p_1 + p_2 = 1$, $\epsilon \sim \mathrm{Gaussian}\,(0, \mathcal{C}_q)$.

2. Accept q' as $q^{*(m+1)}$ with probability

$$\min\left\{1, \exp\left(\ell\left(q', \cdots, \left(\sigma^2\right)^{(m)}\right) - \ell\left(q^{*(m)}, \cdots, \left(\sigma^2\right)^{(m)}\right)\right)\right\}, \tag{3.13}$$

where the likelihood is of form (3.6), and the likelihood ratio can be obtained via

$$\exp\left(-\frac{1}{2\sigma^{2(m)}}\sum_{i=1}^{N}\left(\mathrm{SSE}\left(g_i^{(m+1)}, q'\right) - \mathrm{SSE}\left(g_i^{(m+1)}, q^{*(m)}\right)\right)\right), \tag{3.14}$$

3.2.3 Sampling from Gaussian process

The MCMC algorithm described above needs a specific sampling method from a Gaussian process. Lu proposed that this can be achieved by the Karhunen–Loève (KL) expansion (see [32] for details).

Assume we have a Gaussian process with mean 0 and covariance operator $\mathcal{C}$, and $\mathcal{C}$ can be specified by its eigen pairs $(b_l, \lambda_l^2)_{l=1,\cdots,\infty}$, where $\{b_l\}$ is an orthonormal basis for the function space, and $\sum_{i=l}^{\infty} \lambda_l^2 \leq \infty$. It can be shown that by KL expansion,

$$\xi(t) = \sum_{i=l}^{\infty} e_l b_l(t)$$

is a draw from this Gaussian process, when $\{e_l\}$ are drawn independently from normal distribution $N(0, \lambda_l^2)$. In practice, the infinite sum of $\xi(t)$ is approximated by truncating the sum at a constant $L \in \{1, 2, \cdots\}$.

For the Gaussian prior on g, we use Fourier functions as the eigenfunctions

$$b_l(t) \in \{0, \sqrt{2}\cos(2\pi lt), \sqrt{2}\sin(2\pi lt)\}, \quad l = 1, 2, \cdots, L_g$$

and set eigenvalues to be $\lambda_i^2 = \frac{\sigma_g^2}{i^2}$. For the Gaussian prior on q^*, we use Fourier functions as the eigenfunctions

$$b_i(t) \in \{1, \sqrt{2}\cos(2\pi lt), \sqrt{2}\sin(2\pi lt)\}, \quad l = 1, 2, \cdots, L_{q^*}$$

because $\int_0^1 q^*(t)dt$ is not necessarily 0, and set eigenvalues to be $\lambda_i^2 = \frac{\sigma_{q^*2}}{i^2}$. Both σ_{q^*} and σ_g are assumed to be known.

3.2.4 Standardization

Similar to the discussion in Section 2.2.3, since all functions $f_1, \cdots f_N$ are invariant to a common warping, and the template function q^* is only identifiable up to an equivalence class of warpings, i.e. $[\mu_q] \in \mathcal{L}^2/\Gamma$, a standardization step like the one mentioned at the end of Section 2.2.3, is necessary in the updating process.

The standardization step is appropriate in that shifting the warping functions γ_i and the template q^* by a common $\tilde{\gamma}$ does not change the likelihood [22]. The key invariability that leads to this result is that $(q, \gamma \circ \tilde{\gamma}) = ((q, \gamma), \tilde{\gamma})$ and the $\mathcal{L}_2$ distance between two SRVFs q and q^* is invariant to a common warping.

3.2.5 Recommendations for the hyper-parameters in application

The following quantities need to be specified before applying the Bayesian registration to a data set: (1) number of basis functions L_g used to estimate g, (2) number of basis functions L_q used to estimate the template q^*, (3) σ_g that determines the Gaussian process prior on g, (4) σ_{q^*} that determines the Gaussian process prior on q^*, (5) Inverse Gamma parameter a, b, and (6) grid sizes for both the observations and for the warping functions T_f and T_γ.

As illustrated by Lu in a simulation study regarding pairwise Bayesian registration, $M_g \geq 20$ is sufficient, and based on the deviance information criterion (DIC), a choice of M_g in the range of $20 \sim 30$ is recommended. It is also recommended that the grid sizes T_f and

T_γ should be the same and equal to 200 for better recovery of the observations and warping functions. In addition, M_q is chosen based on the smoothness of the function f, $\sigma_{q^*} = 0.8q_{95}$ (q_{95} is the range of the middle 95% values of all q_i), $\sigma_g = 4, a = b = 0.1$ in the applications presented in [22].

3.2.6 Advantage of Bayesian registration

Apart from the advantages of being indifferent to grid sizes and having intuitive setup to incorporate prior knowledge as mentioned at the end of Section 3.1, we further illustrate the effectiveness of the framework introduced above using an application from the original paper [22].

The height of girls recorded from ages 1 to 18 in the Berkeley growth study [41] is a benchmark dataset used in FDA. The target functional object is the growth rate, represented by the first derivatives of the height curves. Registration is important to this data because individuals arrive at different phases of height growth at different times. The growth rate can be better compared at the same growth window if registered. Omitting discussion on the specific set up of the registration, here we only present the final registration result in Figure 3.2.

In Figure 3.2, the original growth rate functions, their registered versions along with 95% credible interval, SRVF of the original function and the registered functions with 95% credible interval are shown in the top panel; the cross-sectional mean with a two-standard-deviation band before and after registration are shown in the lower panel. Apparently, comparing to the energy-based models which only output the template mean and the warping functions,

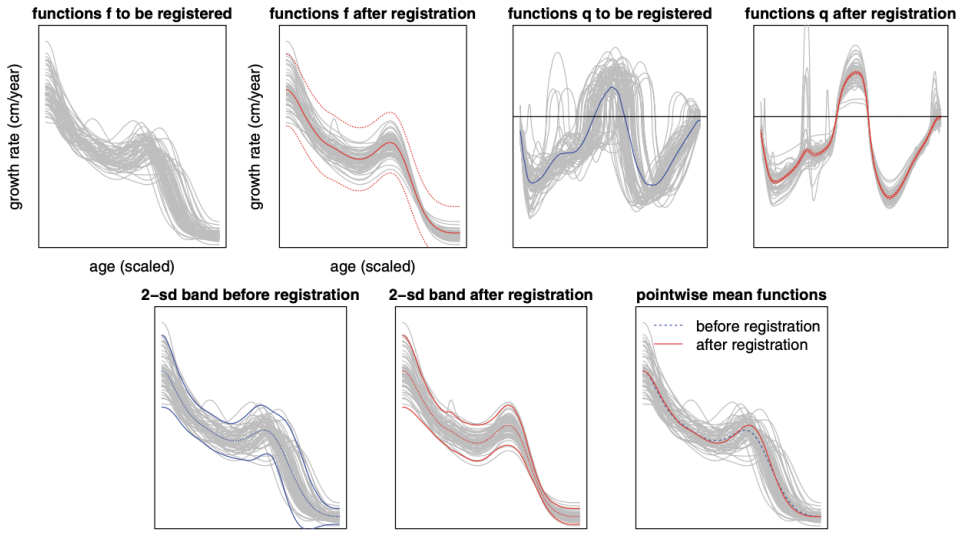


Figure 3.2: Registration results for the Berkeley Growth Study [23]

the Bayesian formulation has the benefit of being able to quantify uncertainty on both the warping functions as well as the template mean, as shown by the credible intervals in this Berkeley Growth study. Apart from that, the Bayesian registration is able to gain insight on the true variability in growth spurt among teenage girls. The mean comparison of the mean and 2-sd band before and after registration shows that: (1) the mean growth rate after registration captures more variability that corresponds to the growth spurt phase and slowing-down phase; (2) the 2-sd band narrows significantly during the post-puberty phase after registration, which picks up the pattern that growth rate varies much less across individuals in this phase.

3.2.7 Disadvantages of Bayesian registration assuming independence

Despite being a relatively efficient Bayesian registration method, this model could render odd estimates of the registered functions if the data have abnormal noises in some of the observed functions.

Here we show one simulated example where we have one sine wave that is distinctly different from the rest of the curves that are to be registered. This exaggerated simulated example is selected merely for the purpose of showcasing the necessity of introducing the dependent priors in the Bayesian registration. It should not be seen as an example that can be related to any practical scenario. Specifics about the simulated data is described in Section 3.4. We show the registration results as well as the original functions and their SRVFs in Figure 3.3.

In Figure 3.3, the original functions, their registered versions along with 95% credible interval, and the estimated warping functions are shown in the top panel; SRVF of the original function and the registered functions with 95% credible interval and the pointwise mean before and after registration are shown in the lower panel. It is clear that the algorithm over-fitted as the sine wave is distorted to match the pattern of the rest of the curves. The estimated warping function of the sine wave (the blue curve in the estimated warping functions) deviates from the identity warping so much, that it could be seen as an outlier among all warping functions.

Even though the prior variance σ_g^2 controls how concentrated the prior distribution of the warping function is around the identity warping $\gamma(t) = t$, this algorithm is still sensitive to

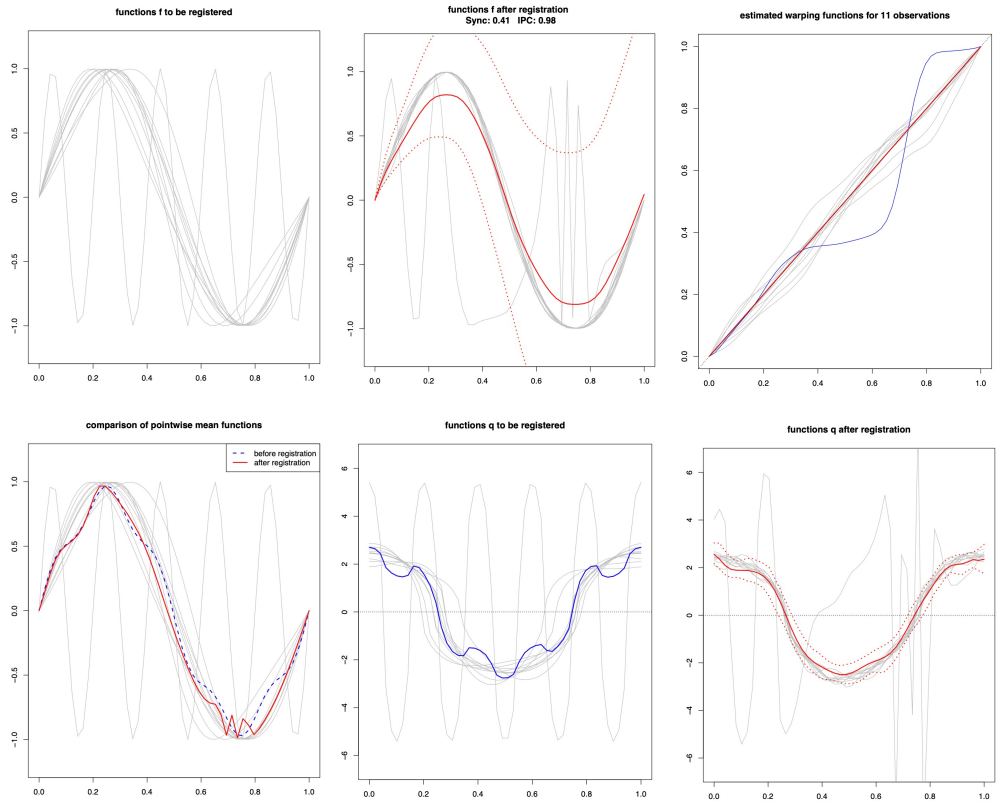


Figure 3.3: An example of over-fitting for the independent Bayesian registration method

the noise in the functions and tends to over-fit. This potential drawback on the independent prior setup also motivates our new method.

3.3 Elastic graph-based function registration

Previously, the prior on $g_1, \cdots, g_N$ is assumed to be independent. However, this is usually not the case for spatially-indexed functional data. In this section, we introduce spatial dependency in the warping priors based on the Bayesian framework mentioned above.

We extend ideas from the conditional autoregressive (CAR) model [3] , which is one of the most popular models in the analysis of spatial data, to the spatially-indexed functional

data, as a way of modeling the spatial dependency. In this setup, geographic regions are considered to be dependent if they share a boundary. This structure can be represented via a finite graph, where an edge represents dependence between two locations.

In the case of fMRI data, voxels are viewed as nodes and two voxels are connected if they are close enough. Using the notation of graph theory, we have $G = (V, E)$, where $V = \{f_1, \cdots, f_N\}$ is a set of nodes, and $E = \{(f_i, f_j), \cdots\}$ is the set of edges. N is the total number of locations where we obtain the functional observations. The connection can be represented via adjacency matrix W where $w_{ij} = 1$ if $(f_i, f_j) \in E$, and $w_{ij} = 0$ otherwise.

The aim here is to allow dependency in the parameter space (i.e. warping functions) based on the spatial structure. Again, we want to work on the tangent space $T_1(\Psi)$ to exploit its linearity property. The following setup is built upon the ideas introduced by Zhang et al. [44].

3.3.1 Prior: functional CAR model

Below, we first introduce the general framework of the functional CAR model introduced by Zhang et al., and then lay out our modification on such model that becomes the prior distribution in our final Bayesian hierarchical model.

separable function CAR

The functional CAR model proposed by Zhang et al. is the functional analog of the classic univariate CAR which uses basis function modeling together with spatial process modeling. In general, the functional CAR model can deal with data generated from a spatial functional process $\{f_i(t) : i \in D\}$ where $D = \{1, \cdots, N\}$ is the index set of a fixed spatial locations

(e.g. from a lattice), and $Y_i(t)$'s are functional random variables that are measured on a common set of T time points.

Instead of focusing on predicting future observations, Zhang et al.'s functional CAR model targets at discovering "highly-local changes within the functional domain", while borrowing strength both spatially and functionally for estimation and inference. The key properties of their model comes from the univariate CAR model. The univariate CAR model can be viewed as a discrete Markov random field in which the full conditional distribution depends on the spatial adjacency. It can model spatially-smooth variances, as well as variables with strong regional correlations.

Suppose $Y_{ij}(t) = \mu_j(t) + E_{ij}(t)$ is a functional response observed in the spatial location A_i in the lattice D for subject $j, j = 1, \cdots, J$ on a set of common timepoints $t = 1, \cdots, T$. In this model, functions from the same subject share the same mean $\mu_j(t)$, and the errors $E_{ij}(t)$ follow a random Gaussian process with mean zero. $Y_{ij}(t)$ can be represented with a truncated basis expansion of form

$$Y_{ij}(t) = \sum_{l=1}^{L} \phi_k(t) y^*_{ijl} = \mathbf{y^*_{ij}} \mathbf{\Phi}(t)$$

where $\mathbf{\Phi}(t)$ is the vector of the basis functions, and $\mathbf{y^*_{ij}}$ is a row vector of the basis coefficient for the (i, j)th curve. Then, in the basis space we have

$$Y^* = X\mu^* + E^*,$$

where Y^*, μ^* and E^* are basis coefficients of the observations, the mean and the error terms respectively.

In a separable function CAR model, it is assumed that in the basis space, the N-dimensional error vector $E^*_{\cdot jl} = \left(E^*_{1jl}, \cdots, E^*_{Njl}\right)'$ for subject j and lth basis functions follows the following CAR prior independently:

$$\mathbf{E}^*_{ijl} \mid \{\mathbf{E}^*_{i'jl}\}_{i' \neq i} \sim N\left(\rho \sum_{i'=1}^{N} \frac{w_{ii'}}{w_{i+}} \mathbf{E}^*_{i'jl}, \quad \frac{s^*_{jl}}{w_{i+}}\right), \tag{3.15}$$

where $w_{ii'}$ is the element of the adjacency matrix W and $w_{i+} = \sum_{i'} w_{ii'}$ is the number of neighbours of the A_i in D.

In this model s^*_{jl} is the scalar parameter that controls the smoothing between connected sites, and ρ dictates the strength of spatial association between such sites.

functional CAR prior for Bayesian registration

We adopt the above separable functional CAR framework for our graph-based registration. Our model is a degenerated case of the separable functional CAR model in that we do not have subject-wise variability. However, since the model assumes independence between subjects, (3.15) can be directly used. The only modification we have is that now the basis-specific smoothing parameters s^*_l's are assumed to be a function of a common smoothing parameter in the form

$$s^*_l = \frac{s}{l^2},$$

in which l depends on the basis element we are looking at. As a result, basis elements with higher frequency have lower prior variance. Apart from that, we also use Fourier basis

functions in our prior setup. Specifically, define Fourier basis for the tangent space $T_1(\Psi)$:

$$\phi_l(t) \in \{0, \sqrt{2}\cos(2\pi lt), \sqrt{2}\sin(2\pi lt)\}, \quad l = 1, 2, \cdots,$$

and $\xi_i = (\xi_{i1}, \cdots, \xi_{iL})$ is the L-dimension representation of $g_i \in T_1(\Psi)$:

$$g_i(t) \approx \sum_{l=1}^{L} \xi_{il}\phi_l(t). \tag{3.16}$$

Then we model spatial dependency in the prior on those Fourier coefficients ξ_i's. Specifically, we assume independent Gaussian distributions among those Fourier basis coefficients with heterogeneous variance-covariance matrices that depend on both basis index l and function index i.

That is, for each $l = 1, \cdots, L$, we allow a form of autoregression given by

$$\xi_{il} \mid \{\xi_{i'l}\}_{i' \neq i} \sim N\left(\rho \sum_{j=1}^{N} \frac{w_{ij}}{w_{i+}}\xi_{jl}, \quad \frac{\sigma_g^2}{l^2 w_{i+}}\right), \quad \text{independent across } l \tag{3.17}$$

where w_{ij} is the element in the adjacency matrix W, $w_{i+} = \sum_j w_{ij}$ represents the number of neighbors of f_i, and $\rho \in (0, 1]$ is the spatial correlation parameter which describes to what extent the neighbors are similar.

It can be easily found that when $\rho \in (0, 1)$, the joint distribution of $\xi_{\cdot l} = (\xi_{1l}, \cdots, \xi_{Nl})$ is a proper multivariate normal distribution:

$$\xi_{\cdot l} \sim MVN(0_N, \frac{\sigma_g^2}{l^2}(D_w - \rho W)^{-1}), \tag{3.18}$$

where $D_w = \text{diag}\{w_{1+}, \cdots, w_{M+}\}$. When $\rho = 1$ in the CAR prior specification, joint prior distribution is not a proper distribution. Even though we do not obtain a proper joint prior distribution, setting ρ =1 is frequently used in practice, and is suitable if the resulting posterior distribution is proper. One example of this is the intrinsic conditional autoregressive (ICAR) model [4].

Based on the K-L expansion properties on Gaussian processes, each function $g_i \in T_1(\Psi)$ is a Gaussian process with the form of Gaussian $(0, \mathcal{C}_{g_i})$.

Next, we explore the properties of the joint prior of $(g_1, \cdots, g_N)$. Let Φ be the $L \times T$-dimensional matrix of Fourier basis functions evaluated at $[t] : (\phi)_{(b,j)} = \phi_b(t_j)$. Then the matrix of discretized functions in the tangent space $T_1(\Psi)$ has the form: $G = (g_m(t_j)) = \Xi\Phi$. Suppose Φ is full rank, which is always true for orthogonal basis system like Fourier basis, then by results from [44], the independent proper CAR prior on each $\xi_{\cdot b}$ induces the following structure on the prior distribution of $\{g_i\}_{i=1}^M$:

(1) At a specific time point t:

$$\{g_i(t)\}_{i=1}^M \sim MVN(0_M, \Omega^{tt})$$

where $\Omega^{tt} = \alpha_t D_w - \beta_t W$;

Here, $\alpha_t = \sum_{l=1}^L \frac{v_{tl}^2 \cdot l^2}{\sigma_g^2}$ and $\beta_t = \sum_{l=1}^L \frac{v_{tl}^2 \cdot l^2}{\sigma_g^2}\rho$ where v_{tb} is an element of the $N \times L$ basis transformation matrix $\mathbb{V} = \Phi'(\Phi\Phi')^{-1}$. Therefore, the spatial correlation at time t is induced by $\frac{\beta_t}{\alpha_t}$.

(2) At different time point t and t', the partial cross-covariance between g_i's have the CAR covariance structure:

$$\Omega_{tt'} = a_{tt'} D_v - b_{tt'} V,$$

where $a_{tt'} = \sum_{l=1}^{L} \frac{v_{tl} v_{t'l} \cdot l^2}{\sigma_g^2}$ and $b_{tt'} = \sum_{l=1}^{L} \frac{v_{tl} v_{t'l} \cdot l^2}{\sigma_g^2} \rho$

Given those properties, we can see that the CAR spatial correlation structure is preserved in $(g_1, \cdots, g_N)$ with the same spatial dependence between any pair of neighbors, even though the basis functions are modeled independently. In summary, the CAR prior structure allows for the model on $(g_1, \cdots, g_N)$ to capture spatial dependence across functions, while also borrowing strength at different time points.

3.3.2 Hierarchical model for spatial function registration

Now that we have specified the transformation on $T_1(\Psi)$ (i.e. the projection on the L-dimensional subspace spanned by the Fourier basis) and an auto-correlation distribution in this space based on the Fourier coefficients, the final hierarchical model can be specified as below.

Model 2:

Prior:

$$\begin{aligned} q^* &\sim \text{ Gaussian } (0, \mathcal{C}_q) \\ \sigma^2 &\sim \text{IG}(\text{ shape } = a, \text{ scale } = b) \\ \xi_{il} \mid \{\xi_{i'l}\}_{i' \neq i} &\sim N\left(\rho \sum_{j=1}^{N} \frac{w_{ij}}{w_{i+}} \xi_{jl}, \quad \frac{\sigma_g^2}{b^2 w_{i+}}\right) \end{aligned} \tag{3.19}$$

where the joint prior is the independent joint distribution of the specified priors above.

Sampling distribution:

$$\{(q_i \circ \gamma_i)([t])\}_{i=1}^N \mid \{\gamma_i\}_{i=1}^N, \sigma^2, q^* \sim \text{(i.i.d.)} MVN(q^*([t]), \sigma^2 I_N) \tag{3.20}$$

where γ_i and ξ_i is linked via g_i (3.5) and a Fourier basis expansion (3.16).

Here, parameters that need to be pre-specified are the number of basis functions L_g (here the footnote g is included to emphasize the basis system is for the space of g_i's), spatial correlation ρ, the variability in the warping functions σ_g^2, Inverse-Gamma parameters a and b, and covariance operator $\mathcal{C}_q$. Using the K-L expansion and assumptions for variance on the basis coefficients in Section 3.2.3, it is sufficient to specify $\mathcal{C}_q$ via L_q and $\sigma_{q^*}^2$.

Note:

ρ can be 0 in *Model 2*. In this case, *Model 2* is equivalent to *Model 1* if G is a fully connected graph. When G is a fully connected graph, i.e., each node is connected to the rest nodes in the graph, (3.20) can be simplified to $\xi_{\cdot l} \sim MVN(0_N, \frac{\sigma_g^2}{l^2 \cdot (N-1) I_N})$. This is exactly the same as the prior of *Model 1* specified Section 3.2.1 when you scale σ_g by $(N-1)$.

3.3.3 MCMC algorithm

Model 2 extends the independent prior structure among $(g_1, \cdots, g_N)$ in *Model 1* to a dependent structure through the basis coefficients. Despite the difference in the prior structure, we can still apply the Metropolis-within-Gibbs (MwG) algorithm introduced for *Model 1* by [23] to the dependent prior case in *Model 2*.

Since the only component changes in the model is the prior on the functions $g_1, \cdots, g_N$, the two models share the same form of full conditional distribution for σ^2 and q^* respectively.

That is, we can still update σ drawing samples from the Inverse-Gamma distribution (3.7) and q^* using the M-H steps with the z-mixture pCN proposal, given the sample and updates of the rest of the parameters.

Here, we lay out the general updating steps for the dependent Bayesian registration algorithm first.

the pCN algorithm for dependent Bayesian function registration

- **Step 1**: Draw $g_i^{(0)}$ from the CAR prior which has $\exp_1^{-1}(g_i^{(0)}) \geq 0$. Set the initial $q^{*(0)}$ to be the point-wise mean of f_i's, and $\sigma^{(0)} = 1$.

For $m = 1, 2, \cdots$, repeat the following steps until the chain has reached the desired chain length.

- **Step 2:** Update $g_1^{(m)}, \cdots, g_N^{(m)}$ to $g_1^{(m+1)}, \cdots, g_N^{(m+1)}$ jointly given $q^{*(m)}$,$\sigma_g^{(m)}$ and $\sigma^{(m)}$:

    1. Make a Z-mixture pCN proposal of the basis coefficients based on the prior CAR mode. Repropose if any of the corresponding g_i has $\exp_1^{-1}(g_i^0) \leq 0$.

    2. Accept the proposal with probability

$$\min\left\{1, \exp\left(\ell\left(g', \cdots, \left(\sigma^2\right)^{(m)}\right) - \ell\left(g_i^{(m)}, \cdots, \left(\sigma^2\right)^{(m)}\right)\right)\right\}, \tag{3.21}$$

where $\ell(\cdot) = \log(L)$, and the likelihood ratio can be simplified to

$$\begin{aligned}\frac{L\left(g_1', \ldots, g_c', q^{*(m)}, \sigma^{2(m)}\right)}{L\left(g_1^{(m)}, \ldots, g_c^{(m)}, q^{*(m)}, \sigma^{2(m)}\right)} &= \frac{\prod_{i=1}^{C} \exp(\frac{\mathrm{SSE_i'}}{2\sigma^{2(m)}})}{\prod_{i=1}^{C} \exp(\frac{\mathrm{SSE_i^m}}{2\sigma^{2(m)}})} \\ &= \exp\left(-\frac{1}{2\sigma^{2(m)}} \sum_{i=1}^{C} \left(\mathrm{SSE}_i' - \mathrm{SSE}_i^{(m)}\right)\right),\end{aligned} \tag{3.22}$$

where $\mathrm{SSE}_i' = \mathrm{SSE}(g_i', q^{*(m)})$ and $\mathrm{SSE}_i^{(m)} = SSE(g_i^{(m)}, q^{*(m)})$. $\mathrm{SSE}(\cdot, \cdot)$ is (3.8).

- **Step 3**: Update $q^{*(m)}$ to $q^{*(m+1)}$ given all the other parameters at their state:

  1. Make a Z-mixture pCN proposal.

  2. Accept q' as $q^{*(m+1)}$ with probability

$$\min\left\{1, \exp\left(\ell\left(q', \cdots, \left(\sigma^2\right)^{(m)}\right) - \ell\left(q^{*(m)}, \cdots, \left(\sigma^2\right)^{(m)}\right)\right)\right\}, \tag{3.23}$$

  where the likelihood is of form (3.6), and the likelihood ratio can be obtained via

$$\exp\left(-\frac{1}{2\sigma^{2(m)}} \sum_{i=1}^{N} \left(\mathrm{SSE}\left(g_i^{(m+1)}, q'\right) - \mathrm{SSE}\left(g_i^{(m+1)}, q^{*(m)}\right)\right)\right), \tag{3.24}$$

- **Step 4**: Transform $g_i^{(m+1)}, i = 1, \cdots, N$ to $\gamma_1^{(m+1)}, i = 1, \cdots, N$, standardize $\gamma_i^{(m+1)}$'s and $q^{*(m+1)}$ using the Karcher mean $\bar{\gamma}$. Finally, transform back the centered γ_i to $g \in T_1(\Psi)$.

- **Step 5**: Update $\sigma^{2(m)}$ to $\sigma^{2(m+1)}$ by drawing a sample from

$$\sigma^{2(m+1)} \sim \text{IG}\left(\text{ shape } = \frac{1}{2}NT + a, \text{ scale } = \frac{1}{2}\sum_{i=1}^{N} \text{SSE}\left(g_i^{(m+1)}, q^{*(m+1)}\right) + b\right). \tag{3.25}$$

Note on the MCMC algorithm:

1. The Z-mixture pCN proposal for updating q^* is the same as the one specified in (3.12).

2. The proposed candidate for $g_1, \cdots, g_N$ jointly is obtained through the proposed updates on the basis coefficients $\xi_{\cdot,l}, l = 1, \cdots, L_g$, which has the prior distribution of form (3.20) given the CAR model assumption (3.19).

   Specifically when adopting the 2-mixture pCN proposal, we propose new $\xi'_{\cdot 1}, \cdots, \xi'_{\cdot L_g}$ iteratively based on

$$\xi'_{\cdot l} = \begin{cases} \xi_{\cdot,l}^{(m)}\sqrt{(1-\beta_1^2)} + \beta_1\epsilon_l & \text{with probability } p_1 \\ \xi_{\cdot,l}^{(m)}\sqrt{(1-\beta_2^2)} + \beta_2\epsilon_l & \text{with probability } p_2 \end{cases} \tag{3.26}$$

   where $p_1 + p_2 = 1$, $\epsilon_l \sim MVN(0_N, \frac{\sigma_g^2}{l^2}(D_w - \rho W)^{-1})$.

3. In practice, the standardization step (i.e. Step 4 of the algorithm) is only performed every 1000 updates of the parameters.

3.4 Simulation study

In this section, we describe a situation where the dependent Bayesian registration method would be useful using a simulated example. The simulated data contains eleven functions

on $[0, 1]$. $f_i(t) = sin(2\pi\gamma_i(t))$ for $i = 1, ..., 10$, where $\gamma_i(t)$ is a random warping function generated from a distribution centered at Karcher mean $\gamma(t) = t$, and the last function is $f_{11}(t) = \sin(10\pi t)$ (see the first plot in Figure 3.6).

In this setting of observed functions, we do not expect the last function (f_{11}) to play a role as important as the rest of the functions in the multiple registration, since it clearly has a different type of shape compared to the other 10 functions. Equivalently, we would like the warping function of f_{11} to be as close to the identity warping as possible. Therefore, the main objective of this simulation study is to investigate if estimated warping of the last sine wave with higher frequency presents any over-fitting issue. To quantify that, we use the Fisher-Rao distance between γ_{id} and $\hat{\gamma}_{11}$ as the criteria (denoted as d_{overfit}).

To use the graph-based registration method, we need to apply a network structure on the simulated data. Specifically, we assume that the eleven functions has the following network structure: the last function f_{11} is connected to the rest of the functions, while the rest of functions are connected to the two functions with the neighouring indices additionally (Figure 3.4).

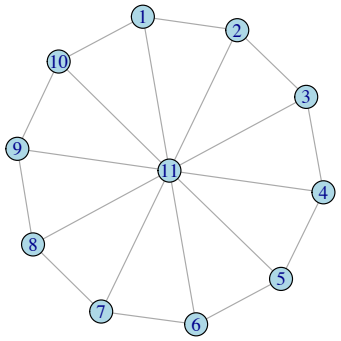


Figure 3.4: The proposed graph structure on the simulated data

This network structure allows the registration of the last function that has much higher frequency than the rest of the sine waves to borrow strength from the other ten functions, so that it will not be over-warped in the multiple registration process.

Now we are ready to apply the elastic graph-based registration method on the eleven functions. The first 20 pairs of the Fourier basis functions are used to estimate the warping functions g_i and the template function q^* (i.e., $L_g = L_{q^*} = 20$). To allow enough spatial dependency among the functions, we set the spatial correlation parameter $\rho = 0.9$. The other constants needed to be specified are chosen according to the recommendations from Lu, Herbei, and Kurtek: $\sigma_g = 4$, $\sigma_{q^*} = 0.8q_{95}$ (q_{95} is defined as the range of the middle 95 values of all q_i), $a = 0.1$ and $b = 0.1$. The 2-mixture pCN proposal is used to propose both the joint of $(g_1, \cdots, g_{11})$ and q^*. The MCMC chain is run for 10^4 iterations with a burn-in period of 0.5×10^4. Trace plot of the log-likelihood (Figure 3.5) shows no indication of lack of convergence.

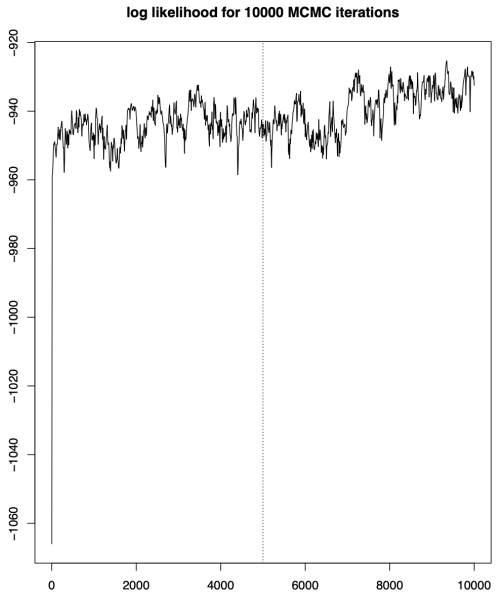


Figure 3.5: Trace plot of the log-likelihood for the simulation study. The horizontal dotted line marks the end of the burn-in step.

The final registration outcomes are shown in Figure 3.6. The graphs are organized in the same with in Figure 3.3: the original growth rate functions, their registered versions along with 95% credible interval, and the estimated warping functions are shown in the top panel; SRVF of the original function and the registered functions with 95% credible interval and the point-wise mean before and after registration are shown in the lower panel.

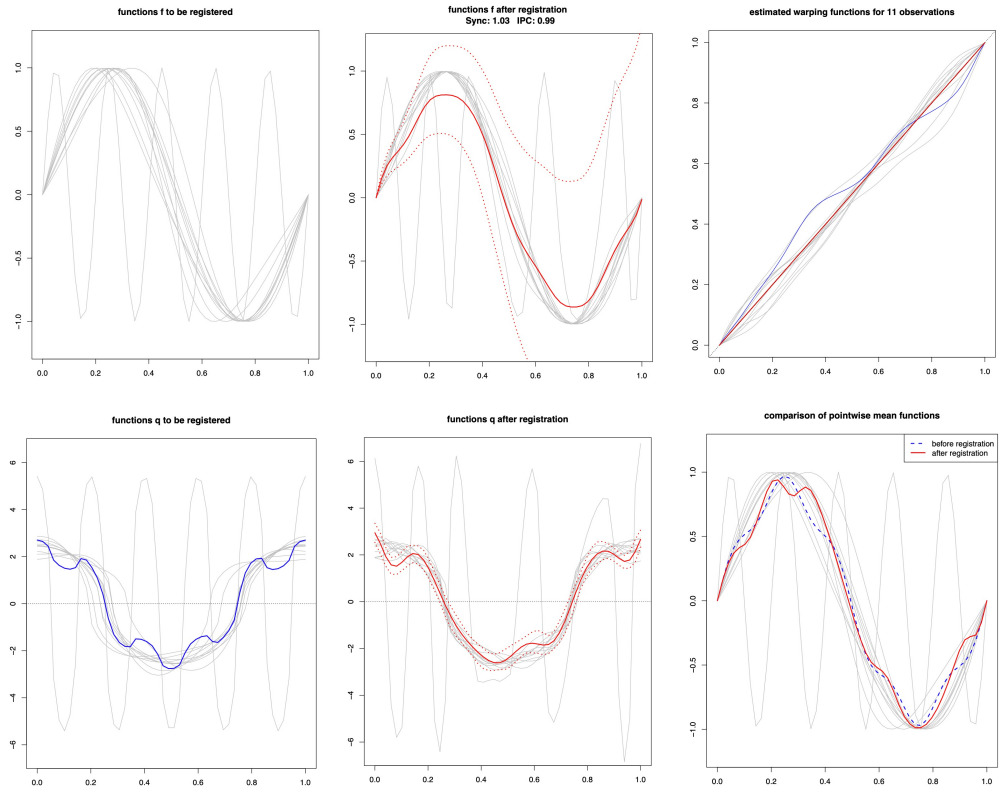


Figure 3.6: A simulation study of the elastic graph-based registration

Comparing our result with the result using Lu et al.'s method, our method has $d_{\text{overfit}} = 0.0403$, while their method has $d_{\text{overfit}} = 0.2787$. Also comparing the registered f_{11}'s, we find that the dependent prior has better performance with respect to over-fitting. The comparison of the registration on f_{11} is shown in Figure 3.7.

The comparison of the point-wise mean after registration is shown in Figure 3.8. We can see that around $t = 0.7$, the point-wise mean after registration with independent priors

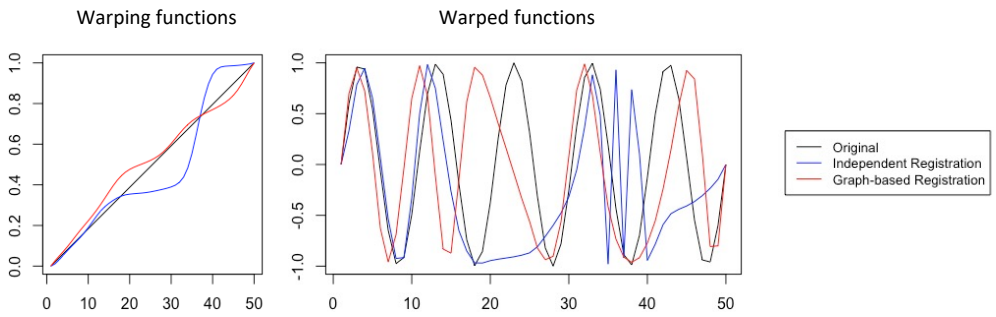


Figure 3.7: Registration comparison on the sine wave with higher frequency

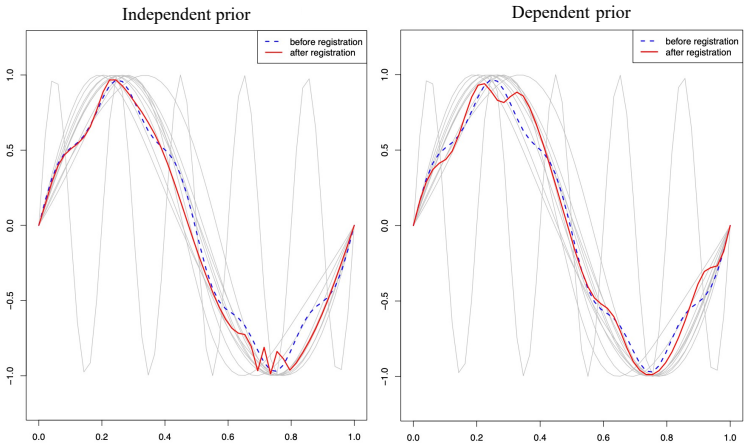


Figure 3.8: Comparison of the registered mean

catches more variability from the noise than that with dependent priors. This is because the previous algorithm "pinched" the last function in this region to fit the trend from the other functions.

Therefore, we confirm in this simulation study that compared to the Bayesian registration using independent priors on g_i's, the graph-based method has successfully prevented the over-fit in the last function, while still managing to match the shapes of the other ten functions.

3.5 A real data application

We also apply the graph-based registration to the fMRI study mentioned in Section 2.5.2. Here we restricted ourselves to one subject from the high cognitive control group and one ROI - the right Medial Frontal Eye Field (rmedFEF). In this case, we have a total of 85 spatially-indexed functions, which have a very low signal-to-noise ratio due to the nature of fMRI scans as mentioned in the overview of fMRI data in Section 2.1. To construct a graph from the given spatial coordinates of the voxels, we assume that two voxels are two nodes connected to each other, if their distance is within the $D_{max} = 8$. Note that D_{max} is also one of the hyper-parameters in the spatial functional SVM algorithm, which defines the neighborhood of the spatial structure. $D_{max} = 8$ is chosen, since this is the most selected neighborhood radius in the cross-validation step in the previous study on this data set. Then we set the spatial correlation parameter $\rho = 0.9$ to allow abundant correlation among the functions. For the MCMC algorithm, the 2-mixture pCN proposal is used to propose candidate functions for both the joint of $(g_1, \cdots, g_{11})$ and q^*, and the MCMC chain is run for 10^4 iterations with a burn-in period of 0.5×10^4.

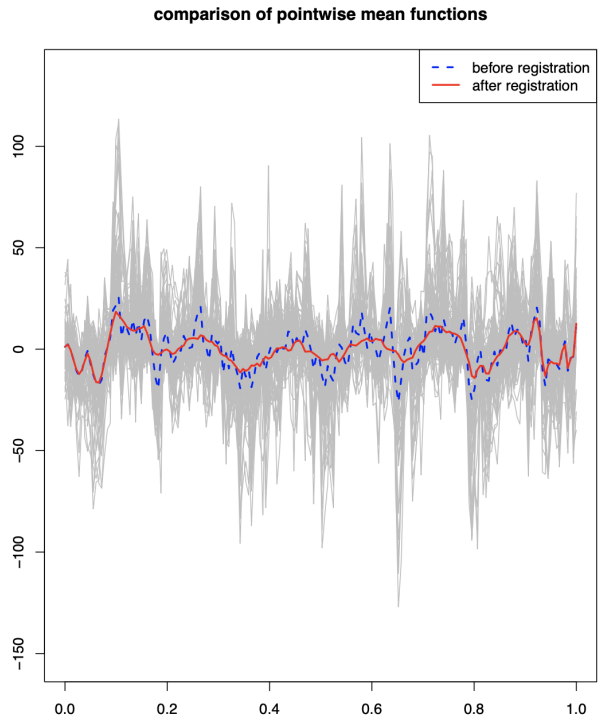
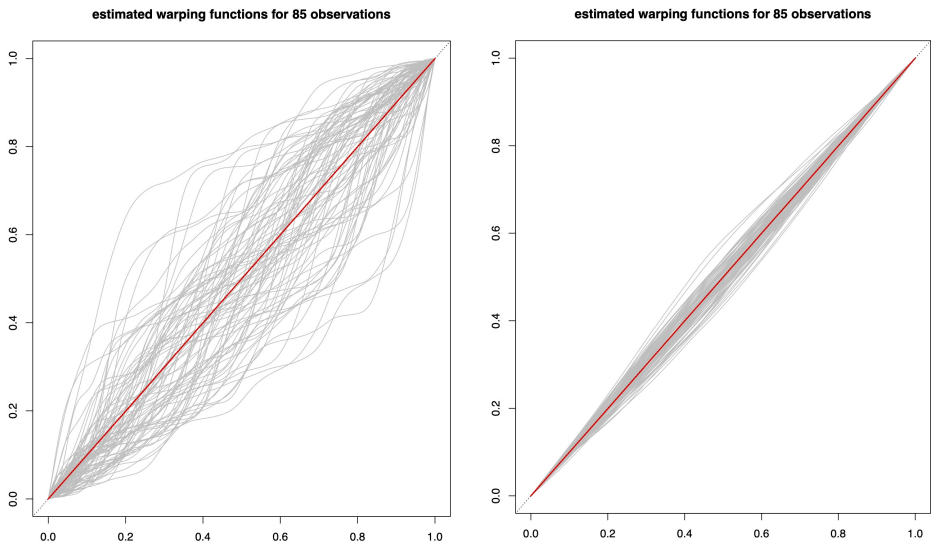


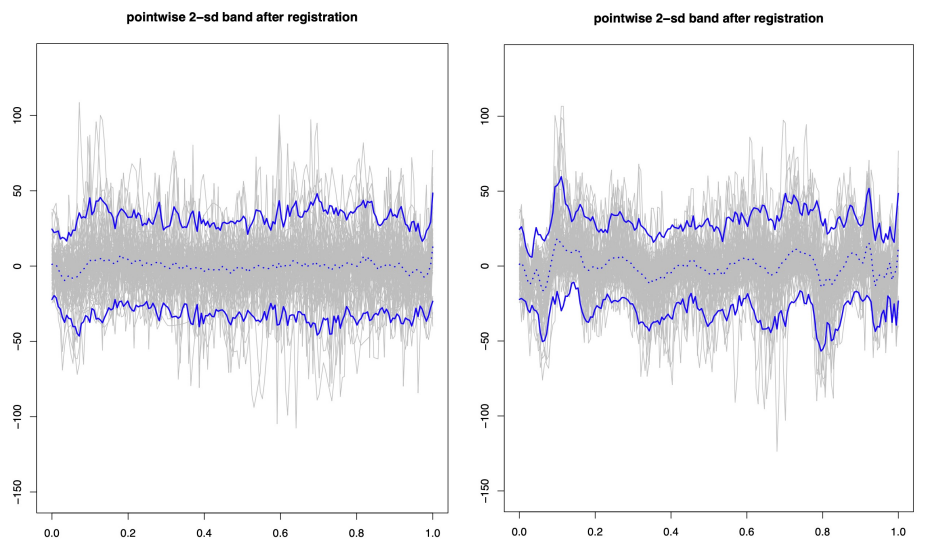
Figure 3.9: An application of the graph-based registration on the fMRI study

The point-wise mean before and after registration is shown in Figure 3.9. The mean after registration contains less noise, and is a better representation of the signals from the block design of this fMRI study. Below we also present the registration results using the Bayesian registration with independent priors on the warpings. The estimated warping functions is presented in Figure 3.10 (left) and the corresponding point-wise mean with a 2-SD band is presented in Figure 3.10. Comparing the point-wise means of the two Bayesian registration models, the i.i.d. model is over-registering the noise (relative to the signal) whereas the spatial dependent model is not registering noise quite as much, thus preserving the block design. This finding is also firmed by comparing the warping functions in both cases: the spatial model has warping functions more close to the identity warping (i.e. $\gamma(t) = t$) than the i.i.d. model.



(a) Independent prior (b) Dependent prior

Figure 3.10: Bayesian registration comparison on the warpings of the fMRI data



(a) Independent prior (b) Dependent prior

Figure 3.11: Bayesian registration comparison on the function means of the fMRI data

3.6 Discussion

3.6.1 Advantage of the graph-based Bayesian registration method

The proposed graph-based Bayesian registration method enjoys the benefits of both using a Bayesian framework and assuming a spatially dependent CAR model.

It has the following features. It (1) allows correlation among the warping functions in the prior; (2) is robust to grid size and scalable to large functional datasets sampled on a fine grid; (3) allows uncertainty analysis of mean amplitude and the warping functions based on the posterior distributions; (4) makes it possible to apply prior knowledge of the smoothness and dependence level to the model.

3.6.2 Future work

In this work, we did not fully investigate how the choices of different network structure as well as different spatial correlation levels specified in ρ affects the posterior estimates. It is of interest that we come up with a sophisticated simulation study that explores those properties of this registration method.

Also, in our method, we only adopted the most simple graph structure, where elements of the edge set E is either zero or one. In a more general network/graph framework, edges can be assigned continuous weights [18]. Graph with edge weights are referred to as "weighted graph". Usually, those weights are assumed to be non-negative by convention, and often are scaled to fall between zero and one. Equivalently, the edge set E can be supplemented with a weighting function which takes inputs as edges and outputs a continuous value, and the

adjacency matrix W has w_{ij} equal to the corresponding weight for the vertex pair i and j. It is possible that we extend our model to include those weighted graphs as the underlying spatial structure for the priors of the warping functions. Additional future directions include considering CAR priors on the SRVFs of the amplitude functions themselves, perhaps with i.i.d. or CAR priors on the warpings.

Chapter 4

Conclusion

In this dissertation, we investigated two specific analytic tasks of spatially-indexed functional data using the framework of elastic data analysis. The proposed methods are based on the work of the multiple registration of functions using the Fisher-Rao metric, which is introduced in Section 1.1 and using an additional Bayesian framework, which is introduced in Section 3.2.

The first task motivated by the fMRI study of schizophrenia aims at classification of each set of spatially-index functional data. The classification tool we adopt is a generalized support vector machine (SVM) algorithm that allows us to adopt any metric of the data space for the adapted kernels in the classifier. In this way, we are able to combine the Fisher-Rao metric as well as spatial weights in the classification algorithm. The final SVM algorithm is specified in Section 2.4.1. We showed the strength of this classification algorithm on simulated fMRI data as well as a real fMRI data set. We also discussed the shortcoming of this algorithm in details in Section 2.6. Apart from those discussions, this algorithm is designed specifically for brain imaging data, resulting in possible lack of generality to data

from other fields. However, in this line of work, we managed to apply the framework of elastic data analysis to the SVM algorithm, which is a new work in the area of classification models in this framework. We hope that this work provides an additional classification tool in this framework that uses the Fisher-Rao metric on a Riemannian manifold.

The second task aims at expanding the Bayesian registration methods that assumes independence among the functions so that they can be applied to spatially-indexed functions. We chose the Bayesian framework that models priors of the warping functions on the tangent space, because we could apply the well defined conditional autoregressive models in this linear space. The spatial dependency is introduced via a functional CAR model which is specified in Section 3.3.1 in the space of the warping functions. The general model structure is introduced in Section 3.3.2 and the MCMC algorithm is explained in Section 3.3.3. We showed how this registration algorithm could borrow strength from neighbouring functions to estimate the warping functions in a simulation study in Section 3.4. We also discussed how this framework is readily available for more general graphs like weighted graphs in Section 3.6. In this model, we are able to extend the independent prior based Bayesian registration to a dependent prior, while still maintaining the model's robustness to grid sizes. This proposed method requires further investigation in its performance with different choices of hyper-parameters, but its framework makes it ready to be applied to the case where we have randomly sampled sets of spatially-indexed functional data. From there, this Bayes model can be combined with the general functional mixed effect model framework which specifies different experimental designs as well as related covariates.

Bibliography

[1] Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[2] R. Bellman and R. Kalaba. "On adaptive control processes". In: *IRE Transactions on Automatic Control* 4.2 (1959), pp. 1–9. DOI: 10.1109/TAC.1959.1104847.

[3] Julian Besag. "Spatial interaction and the statistical analysis of lattice systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 192–225.

[4] Julian Besag and Charles Kooperberg. "On conditional and intrinsic autoregressions". In: *Biometrika* 82.4 (1995), pp. 733–746.

[5] Bharat B Biswal. "Resting state fMRI: a personal history". In: *Neuroimage* 62.2 (2012), pp. 938–944.

[6] Imen Boulnemour and Bachir Boucheham. "QP-DTW: Upgrading Dynamic Time Warping to handle Quasi Periodic Time Series Alignment". In: *Journal of Information Processing Systems* (Oct. 2017). DOI: 10.3745/JIPS.

[7] Wen Cheng, Ian L. Dryden, and Xianzheng Huang. "Bayesian Registration of Functions and Curves". In: *Bayesian Analysis* 11.2 (2016), pp. 447–475. DOI: 10.1214/15-BA957. URL: https://doi.org/10.1214/15-BA957.

[8] Siddhartha Chib and Edward Greenberg. "Understanding the metropolis-hastings algorithm". In: *The american statistician* 49.4 (1995), pp. 327–335.

[9] Kara A Dyckman et al. "An effect of context on saccade-related behavior and brain activity". In: *Neuroimage* 36.3 (2007), pp. 774–784.

[10] Ani Eloyan et al. "Analytic Programming with fMRI Data: A Quick-Start Guide for Statisticians Using R". In: *PLOS ONE* 9.2 (Feb. 2014), pp. 1–13. DOI: 10.1371/journal.pone.0089470. URL: https://doi.org/10.1371/journal.pone.0089470.

[11] "Functional magnetic resonance imaging in schizophrenia". In: *Schizophrenia* 12.3 (Sept. 2010), pp. 333–343. DOI: 10.31887/dcns.2010.12.3/rgur. URL: https://doi.org/10.31887%2Fdcns.2010.12.3%2Frgur.

[12] A. Gelman et al. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN: 9781439840955.

[13] Daniel Gervini and Theo Gasser. "Self-modelling warping functions". In: *Journal of the Royal Statistical Society Series B* 66 (Nov. 2004), pp. 959–971. DOI: 10.1111/j.1467-9868.2004.B5582.x.

[14] Gadi Goelman et al. "Frequency-phase analysis of resting-state functional MRI". In: *Scientific reports* 7.1 (2017), pp. 1–13.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learnin*. 2009.

[16] MR Haylock et al. "A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006". In: *Journal of Geophysical Research: Atmospheres* 113.D20 (2008).

[17] D.G. Kendall et al. *Shape and Shape Theory*. Wiley, London, 1999.

[18] Eric D Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*. Vol. 65. Springer, 2014.

[19] Nicole A Lazar. *The statistical analysis of functional MRI data*. Vol. 7. Springer, 2008.

[20] Denis Le Bihan et al. "Activation of human primary visual cortex during visual recall: a magnetic resonance imaging study". In: *Proceedings of the National Academy of Sciences* 90.24 (1993), pp. 11802–11805.

[21] David S. Lee et al. "Measuring Brain Connectivity via Shape Analysis of fMRI Time Courses and Spectra". In: *Connectomics in NeuroImaging*. Springer International Publishing, 2017, pp. 125–133. DOI: 10.1007/978-3-319-67159-8_15. URL: https://doi.org/10.1007%2F978-3-319-67159-8_15.

[22] Yi Lu. "Function Registration from a Bayesian Perspective". PhD thesis. The Ohio State University, 2017.

[23] Yi Lu, Radu Herbei, and Sebastian Kurtek. "Bayesian Registration of Functions With a Gaussian Process Prior". In: *Journal of Computational and Graphical Statistics* 26.4 (2017), pp. 894–904. DOI: 10.1080/10618600.2017.1336444. eprint: https://

doi.org/10.1080/10618600.2017.1336444. URL: https://doi.org/10.1080/10618600.2017.1336444.

[24] Tadeusz Majewski et al. "Understanding the development of human bladder cancer by using a whole-organ genomic mapping strategy". In: *Laboratory investigation* 88.7 (2008), pp. 694–721.

[25] C. Myers, L. Rabiner, and A. Rosenberg. "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.6 (1980), pp. 623–635. DOI: 10.1109/TASSP.1980.1163491.

[26] Russell A Poldrack. "Region of interest analysis for fMRI". In: *Social cognitive and affective neuroscience* 2.1 (2007), pp. 67–70.

[27] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. 2005.

[28] James O Ramsay. "When the data are functions". In: *Psychometrika* 47.4 (1982), pp. 379–396.

[29] James O Ramsay and Xiaochun Li. "Curve registration". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2 (1998), pp. 351–363.

[30] Fabrice Rossi and Nathalie Villa. "Support vector machine for functional data classification". In: *Neurocomputing* 69.7-9 (2006), pp. 730–742.

[31] Arunava Samaddar. "Statistical learning methods for group analysis in fMRI data". PhD thesis. University of Georgia, 2019.

[32] D.N. Shanbhag et al. *Stochastic Processes: Theory and Methods*. Handbook of Statistics. Elsevier Science, 2001. ISBN: 9780444500144.

[33] Anuj Srivastava and Eric P. Klassen. *Functional and Shape Data Analysis*. Springer New York, 2016.

[34] Anuj Srivastava et al. "Registration of functional data using Fisher-Rao metric". In: *arXiv preprint arXiv:1103.3817* (2011).

[35] Luca Steardo et al. "Application of Support Vector Machine on fMRI Data as Biomarkers in Schizophrenia Diagnosis: A Systematic Review". In: *Frontiers in Psychiatry* 11 (2020), p. 588. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2020.00588. URL: https://www.frontiersin.org/article/10.3389/fpsyt.2020.00588.

[36] Donatello Telesca and Lurdes Y T Inoue. "Bayesian hierarchical curve registration". In: *Journal of the American Statistical Association* 103.481 (2008), pp. 328–339.

[37] Tian Siva Tian. "Functional data analysis in brain imaging studies". In: *Frontiers in psychology* 1 (2010), p. 35.

[38] J. Derek Tucker. *fdasrvf: Elastic Functional Data Analysis*. R package version 1.9.4. 2020. URL: https://CRAN.R-project.org/package=fdasrvf.

[39] James Derek Tucker, John R. Lewis, and Anuj Srivastava. "Elastic functional principal component regression". In: *Statistical Analysis and Data Mining* 12.2 (Dec. 2018). ISSN: 1932-1864. DOI: 10.1002/sam.11399. URL: https://www.osti.gov/biblio/1492372.

[40] James Derek Tucker, Wei Wu, and Anuj Srivastava. "Generative models for functional data using phase and amplitude separation". In: *Comput. Stat. Data Anal.* 61 (2013), pp. 50–66.

[41] Read D Tuddenham. "Physical growth of California boys and girls from birth to eighteen years". In: *University of California publications in child development* 1 (1954), pp. 183–364.

[42] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. "Functional Data Analysis". In: *Annual Review of Statistics and Its Application* 3.1 (2016), pp. 257–295. DOI: 10.1146/annurev-statistics-041715-033624. eprint: https://doi.org/10.1146/annurev-statistics-041715-033624. URL: https://doi.org/10.1146/annurev-statistics-041715-033624.

[43] Marijke Welvaert et al. "neuRosim: An R package for generating fMRI data". In: *Journal of Statistical Software* 44.10 (2011), pp. 1–18.

[44] Lin Zhang et al. "Functional CAR models for large spatially correlated functional datasets". In: *Journal of the American Statistical Association* 111.514 (2016), pp. 772–786. ISSN: 0162-1459. DOI: 10.1080/01621459.2015.1042581. URL: https://europepmc.org/articles/PMC5176110.