Topics in Functional Data Analysis

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Abhirup Mallik

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

Snigdhansu Chatterjee

June, 2017

Acknowledgements

First, I would like to thank my advisor Prof. Snigdhansu Chatterjee for his continuous guidance and support throughout my graduate life. I have learned a lot about research methods and technical mentorship from him over the years.

I would also like to thank my committee members Prof. Adam Rothman, Prof. Arindam Banerjee, Prof. Nathaniel Helwig for always being open to discussions and all of their constructive comments and suggestions.

I must thank all other faculty members, including Prof. Dennis Cook, Prof. Galin Jones, Prof. Charles Geyer, Prof. Lan Wang, Prof. Sanford Weisberg, who I have interacted with or taken courses from, as many of my learnings from them have helped shape the direction of my dissertation.

A huge thanks goes to Prof. Zack Almquist and Army Research Office for generously supporting my research for the academic years 2015-17. My dissertation was partially supported by ARO YIP Award # W911NF-14-1-0577.

I would also like to thank Dr. Aaron Rendahl from the Consulting Center and the climate research group for funding me at various times.

I want to thank all my managers and mentors during my summer internship

at Google and Travelers insurance for providing invaluable insights into industrial applications of statistics.

I am grateful to my wonderful friends and colleagues at the school of statistics for their unwavering spirit of exploration, competition and cooperation. I consider myself extremely fortunate to be among such a talented and encouraging cohort of friends.

I should also thank the great staff at the office of the school of statistics for helping me sort out administrative issues numerous times.

Finally it goes without saying, I want to thank my parents and my sister for their unconditional love and support.

Dedication

To my parents

Contents

List of Tables

List of Figures

Chapter 1

Review and Introduction to Functional Data Analysis

1.1 Introduction

We begin this document with a brief discussion of functional data as we will be focusing on several applications on functional data analysis later. Even though several applications discussed can be generalized in several other types of data, in our examples we try to focus on methodologies tailored to functional data analysis. So, here we will consider the required backgrounds and definitions that we will need later.

As we move towards an increasingly connected world, the amount of data generated each day has increased by many fold over a short period of time. The majority of the data in the world is estimated to have been generated within past few years, which is relatively short time period considering the history of information sharing. With the advancement of data collection technology including sensors, it is much easier to collect data continuously over long periods of time. These kinds of data are prevalent in diverse fields of study including finance, medical imaging, engineering, climate informatics, chemometrics, etc. In most of these cases we record observations over a period of time and sometimes over multiple locations. With the growth in technologies for internet of things, it is only expected that this kind of data collection will increase in future as well. These kind of data have a common characteristic, these are observations relative to an independently varying parameter.

This kind of data, where the observations are functions of an independently varying parameter is called functional data. If an observation is recorded from time $(t_{\min}, t_{\max})$, then it can be represented by a family of random variables $\{\mathcal{X}(t_j)\}$.

Where $\mathcal{X}(t)$ is a function of independently varying parameter t. A random variable $\mathcal{X}$ is called a functional random variable if it takes values in a functional space. An instance of the functional random variable will be called as functional data. And a data set $\mathcal{X}_1, \cdots, \mathcal{X}_N$ containing N functional random variables is a functional data set. Here we note that the random variable $\mathcal{X}$ being a function must have a defined support, identified by $\mathcal{X} = \{\mathcal{X}(t); t \in T\}$.

We also note that the independent parameter denoted by t, which is most easily imagined as time, can be any other parameter. We will in our examples will see examples of parameters other than time. For example in a chemometric study on decay of weights of materials under applied heat would record the mass loss curve with temperature. In a data set from a spectrometer, response is a function of wavelengths. For the case of a spatio temporal data set, we can consider the data as a function of both time and location. Thus the support set T can be from uni dimensional or multidimensional set of positive reals.

Functional data is very common in engineering disciplines including study of speech. The phoneme data set is described in Elements of Statistical Learning Hastie et al. [3]. It provides a labeled set of curves of log periodograms with five classes. Here we show a sample of the curves with out the label information in figure 1.1. This is an example of functional data with independent parameter as wavelengths.

In climate science data sets of curves are regularly collected and analyzed. Here, in figure 1.2 we are showing the NOx levels measured in a control station in Barcelona over a period of 24 hours in a day. As the data set is for consecutive

days there are clear correlation among the consecutive curves. These are issues we will have to consider while analyzing such data sets.

An image can be considered as a locus of points which are functions of the respective coordinates. In figure 1.3 we show the trace of images as functions of their horizontal component. Interestingly, these images were writing of the word "fda" on paper by Ramsey and this data set is referenced in their book Ramsey J and Silverman B W [4] on functional data analysis.

It is to be noted that even though these kind of data is theoretically conceptualized as functions of an independently varying parameter, in reality, the observed curve is recorded over a grid of points. Depending on how dense this observation grid is, the data set might need to be smoothed. The smoothness assumption is necessary in functional data analysis as we will describe later, these functions are often modeled using basis functions, which are applicable on smooth functions only.

So, in summary, functional data analysis deals with data sets where each observation is considered to be a function. These kinds of data sets are naturally occurring in various fields of engineering, climate, finance, econometric, etc. The term functional data analysis was first coined by Ramsey, who later published a standard text Ramsey J and Silverman B W [4] in this literature. However, application of functional analysis on data analysis goes back to Greander and Rao in fifties. As these kind of data are theoretically infinite dimensional, and the observed data intrinsically high dimensional before smoothing, this gives rise to various approaches in analyzing this kind of data as well as exploring theoretical

properties of the functions.

1.2 Representation of Functional Data

As the functional or curve data sets are real valued functions $X_1(t), \cdots, X_n(t)$ over a compact interval $t \in T$, these are usually viewed as realizations of stochastic processes, often assumed to be in a Hilbert space such as $L^2[T]$. A stochastic process $X(t)$ is called $L^2[T]$ if $E(\int_T X^2 dt) < \infty$. It is also possible to view these kind of data as high dimensional data and model them with parametric approaches with assumptions of some internal structure like sparsity. However, such approaches face the challenge of mis-specified assumptions and natural ordering of independent axis. Non and semi parametric approaches are most popular choice for dealing with functional data, as they offer flexibility as well as take advantage of the smoothness properties of the data.

In terms of complexity, Wang et al. [5] in their review paper broadly classifies functional data into two categories based on simplicity of their origin, first generation and next generation. First generation functional data set is a set of curves, most common if a single data set is being considered. The examples provided in this chapter are all of this nature. The next generation functional data is a part of some other complex data objects and they involve deeper analysis problems. These data may be correlated or be repeated measurements or might have some other kind of structural element in them. The learning problems that we consider later in this document are of this nature.

As the observed functional data is recorded on a grid, which might or might not be pre defined, we should consider the spacing among the points in the grid. If the data is originated from an instrument or a sensor, which records observations in a fixed interval, then we will find a data set on a equally spaced grid. Such data will be called balanced functional data. However, in many experimental situation, the grid points are not pre determined, and data is recorded irregularly. These kind of data, still being functional in nature will be called imbalanced functional data. In our work, we will not go into details of such data sets, although we do acknowledge that imbalanced data sets are quite common in nature. All the applications considered in this document have equally spaced grid.

The spacing between grid points $t_j - t_{j-1}$ determine how densely the data has been sampled. As most often functional data is smoothed, it is helpful to have more data for better smoothing. So, high dimensional nature of the data might cause problems in traditional method of analysis, but here, it actually helps to have a denser grid. In fact there has been some theoretical advancement on trying to quantify the border when functional approaches have more advantage over traditional analysis techniques based on the density of the observations. Formal definition of dense functional data is still lacking in this literature. Usually, if the estimates can achieve a $\sqrt{n}$ consistency when the number of points in the observation grid converges to infinity, the data is called dense. We refer to Zhang and Wang [6] for further details on this approach. In our work, we will not enter into this classification of dense or non dense data sets, as we will only stick to the functional approaches.

As most of our analysis will be on the observed functional data, we need to specify how we plan to measure distances between them. We follow the semi parametric approach of Ferraty [2], where they have conceptualized the observed functional data in a semi normed space. This is where the functional approach starts to differ from the traditional multivariate analysis approach. As using Euclidean or any other point wise norm in this case will result in too much noise, as what we are interested in are distances between two functions. While, this would work in theory, but in practice, using point wise norms will magnify the errors in each point as thus resulting in poor estimates. We consider the observed functional data in a semi normed space. We recall some definitions as

Definition 1. $||.||$ *is a semi norm on a functional space* $\mathcal{F}$ *if*

1. $\forall(\lambda, x) \in \mathbb{R} \times \mathcal{F}$, $||\lambda x|| = ||\lambda||||x||$

2. $\forall(x, y) \in \mathcal{F} \times \mathcal{F}$, $||x + y|| \leq ||x|| + ||y||$

The difference between semi norm and norm is that $||x|| = 0$ does not imply $x = 0$. This is used to define semi metric $d(,)$ as

Definition 2. $d(.,.)$ *is a semi metric on a functional space* $\mathcal{F}$ *if*

1. $\forall x \in \mathcal{F}$, $d(x, x) = 0$

2. *Triangle inequality,* $\forall(x, y, z) \in \mathcal{F} \times \mathcal{F} \times \mathcal{F}$, $d(x, y) \leq d(x, z) + d(y, z)$

Thus relaxing on the uniqueness property allows us to use projections of the data that are approximate to define distances. These projections are used to

generate a lower dimensional representation of the data. As observed functional data is inherently high dimensional, we needed to use some form of dimension reduction technique to deal with this data. Functional data analysis approaches this by incorporating the dimension reduction into the definition of the distances.

The dimension reduction methods serve two purposes here. Obviously they are used to define semi metrics that are in much smaller dimension than the original observed data. Besides, many of the dimension reduction techniques depend on basis representation, which often have physical interpretations. Many properties about the underlying stochastic process can be inferred by analyzing the coefficients of the basis representation.

1.3 Examples of Semi metrics

One of the most popular method for dimension reduction is principal component analysis. It is widely used in multivariate data analysis for wide range of applications. In the domain of functional data analysis, it is termed as functional principal component analysis. The idea of using principal components is quite old and goes back to fifties. In one of our later applications, we explore the origins and history of the principal components in functional data and climate applications in detail. FPCA decomposes the observed functional data into a sequence of orthogonal random variables. For practical applications, only a finite number of those components are used to describe the process. The components then can be used for usual multivariate analysis, thus principal components acting as a way to

leverage multivariate statistical techniques in the functional world.

The definition of semi metric based on principal component decomposition is quite trivial. We first formally define the mean and covariance operator for the functional data. For functional random variable $\mathcal{X}(t) \in \mathcal{F}, t \in T$, the mean function $\mu(t) = E(\mathcal{X}(t))$. The covariance function for two functions (s,t) in L^2 is given by

$$\Gamma_{\mathcal{X}}(s,t) = E(\mathcal{X}(s)\mathcal{X}(t)) \tag{1.1}$$

Following the terminology of functional analysis (Conway [1]), this is a compact Hilbert-Schmidt operator. As this is non negative definite, it should have non negative real eigenvalues $lambda_1 \geq \lambda_2 \geq \cdots$. With the application of Mercers theorem, using spectral decomposition of $\Sigma(,)$

$$\mathcal{X} = \sum_{k=1}^{\infty} \Big(\int \mathcal{X}(t)\nu_k(t)dt \Big)\nu_k \tag{1.2}$$

Here, we can limit the number of components in the above decomposition to approximate the function with finite number of coefficients. If we choose $q > 0$ components, the approximate function can be written as

$$\mathcal{X}^{(q)} = \sum_{k=1}^{q} \Big(\int \mathcal{X}(t)\nu_k(t)dt \Big)\nu_k \tag{1.3}$$

The convergence of the equation (4.2) holds uniformly such that $\sup_{t\in T} E[\mathcal{X}(t) - \mathcal{X}^{(q)}(t)] \to 0$ as $q \to \infty$. Thus the equation (4.3) provides a good approximation of the process with finite number of components. Thus the information in $\mathcal{X}$ is summarized by a q dimensional vector $\mathcal{X}(t)\nu_k(t)$. This dimension reduction can

be used to construct a semi metric as follows

$$d_q^{PC}(\mathcal{X}_i, \mathcal{X}) = \sqrt{\sum_{k=1}^{q} \Big(\int (\mathcal{X}_i(t) - \mathcal{X}(t))\nu_k(t)dt \Big)^2} \tag{1.4}$$

This theoretical definition can be used to calculate the principal component semi metric in practice. The covariance operator $\Gamma_{\mathcal{X}}(s,t)$ is not observed, so must be estimated. We use the empirical version to estimate this.

$$\hat{\Gamma}_{\mathcal{X}}(s,t) = \sum_{i=1}^{n} (\mathcal{X}_i(s)\mathcal{X}_i(t))/n \tag{1.5}$$

Similarly, we also approximate the integral in equation (4.4) with sum, thus approximating the principal component semi metric by its empirical version.

$$\hat{d}_q^{PC}(\mathcal{X}_i, \mathcal{X}) = \sqrt{\sum_{k=1}^{q} \Big(\sum_{j=1}^{n} (\mathcal{X}_i(t) - \mathcal{X}(t))\nu_k(t)_j \Big)^2} \tag{1.6}$$

Computation for this semi metric is quite easy as in most computational platform, principal component analysis using eigen decomposition is implemented, which can easily be used for this calculation.

There are several other popular semi metrics available in the functional data analysis literature. In the family of basis function reorientation can also be used in this context. If we denote the basis functions by $\phi_k(t)$, then they cay be used for expansion as

$$\mathcal{X}(t) \approx \sum_{k=1}^{q} \beta_k \phi_k(t) \tag{1.7}$$

For Fourier basis family, $\phi_k(t) = a_k \cos(\frac{2\pi nx}{T} + b_k \sin(\frac{2\pi nx}{T}))$, where the Fourier coefficients (a_n, b_n) are given by $a_n = \frac{2}{T}\int_0^T \mathcal{X} \cos(\frac{2\pi nx}{T})$ and $b_n = \frac{2}{T}\int_0^T \mathcal{X} \sin(\frac{2\pi nx}{T})$

Similarly, B-spline family of basis functions are also equally popular for their flexibility for modeling with real data sets. In figure 1.4 two basis functions are shown for small number of basis. With high enough number of basis functions, we can use linear combinations of these functions to approximate any functional data set.

The use of these basis families in constructing semi metrics is by their coefficient representations. As all functional basis families can be used to take a linear combination of appropriate of appropriate order to approximate an observed functional data, the coefficients will be considered as the representation of the functional data in the basis space and the semi metric between two functional data points will be calculated by the distance between their coefficients in the same basis space.

Another common approach of constructing semi metrics using parametric basis families is using distances among their derivatives. If a functional data is approximated with basis family $\mathcal{B}$ as (1.8)

$$\hat{\mathcal{X}}^{(q)}(t) = \sum_{b=1}^{B} \hat{\beta}_b B_b(t) \tag{1.8}$$

As the family of the basis functions are well known, getting derivatives is simpler. The representation of the observed data in terms of basis function would be closer to each other if a finite number of their derivatives are also close. Hence, the derivative semi metric can be defined as

$$d_q^{deriv}(\mathcal{X}_i, \mathcal{X}) = \sqrt{\sum_{k=1}^{K} w_k (\sum_{j=1}^{n} (\hat{\mathcal{X}}_i^{(q)}(t) - \hat{\mathcal{X}}^{(q)}(t)) \nu_k(t)_j)^2} \tag{1.9}$$

The numerical approximation of the integral in the calculation of the semi metric is denoted by w_k, this can be done using various methods like the Gauss method or area method.

The derivative distance is dependent on the basis family chosen to represent the data, so, with changes in the order of derivative and the basis family, there could be various variants of this distance. It is also common to consider the sum of the distances using derivatives up to a certain order. In figure 1.5, the distances of the Tecator data set is shown using Fourier and B spline basis family representation and using derivative of order 1 and 2.

1.4 Some Frequently Used Datasets

Here we describe some frequently used data sets that are referred later in this document. Most of these data sets are available as a part of the R packages *fda.usc* or *fda*.

1.4.1 Tecator Data

This is a spectrometric data set for meat samples. This data is collected using Tecator Infratec Food and Feed Analyzer using wavelength of 850 - 1050 nm by the Near Infrared Transmission principle. Each sample contains chopped meats and the observations were on protein and fat contents. This is documented in *fda.usc*

and *caret* package. We are using the absorbane data only from the 215 samples. The absorbance is log 10 of the transmittance measured by the spectrometer. The protein and fat contents are measured using laboratory methods. We have used this data set to demonstrate various semi metrics and other methods on functional data analysis.

1.4.2 Phoneme Data

This is originally documented in `http://statweb.stanford.edu/~tibs/ElemStatLearn/`, where it is described to be collected by Andreas Buja, Werner Stuezle and Martin Maechler. Five phonemes were selected for classification, however, we mostly have used it for two class problems or illustration purposes only. The subset we have used is available from `http://www.math.univ-toulouse.fr/staph/npfda/npfda-phoneme.dat`, where 2000 pairs of $(\mathcal{X}, y)$ were noted. The $\mathcal{X}$ represent the discretized log-periodograms and y are the class information. Out of 250 time points in the original data, here we are using the 150 time points selected by the *fda* package.

1.4.3 Canadian Weather Data

This data set is taken from the R package *fda* and also documented in the book by Ramsey J and Silverman B W [4]. It consists of daily temperature and precipitation of 35 different locations in Canada from 1960 to 1994. The daily observations averaged over all the years are used as a time series of weather data from multiple

locations. The data also notes the coordinates of the corresponding weather stations. We have used this data in some illustrations of our semi metrics definition in chapter 4.

Bibliography

[1] John B. Conway. *A course in functional analysis*. Number 96 in Graduate texts in mathematics. Springer, New York, 2nd ed edition, 1997. ISBN 978-0-387-97245-9.

[2] Vieu Philippe Ferraty, FrÃl'dÃl'ric. *Nonparametric Functional Data Analysis*. 2006.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[4] Ramsey J and Silverman B W. *Functional Data Analysis*. 2005.

[5] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg MÃijller. Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, June 2016. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-041715-033624. URL `http://www.annualreviews.org/doi/10.1146/annurev-statistics-041715-033624`.

[6] Xiaoke Zhang and Jane-Ling Wang. From Sparse to Dense Functional Data and Beyond. *Annals of Statistics (Submitted)*.

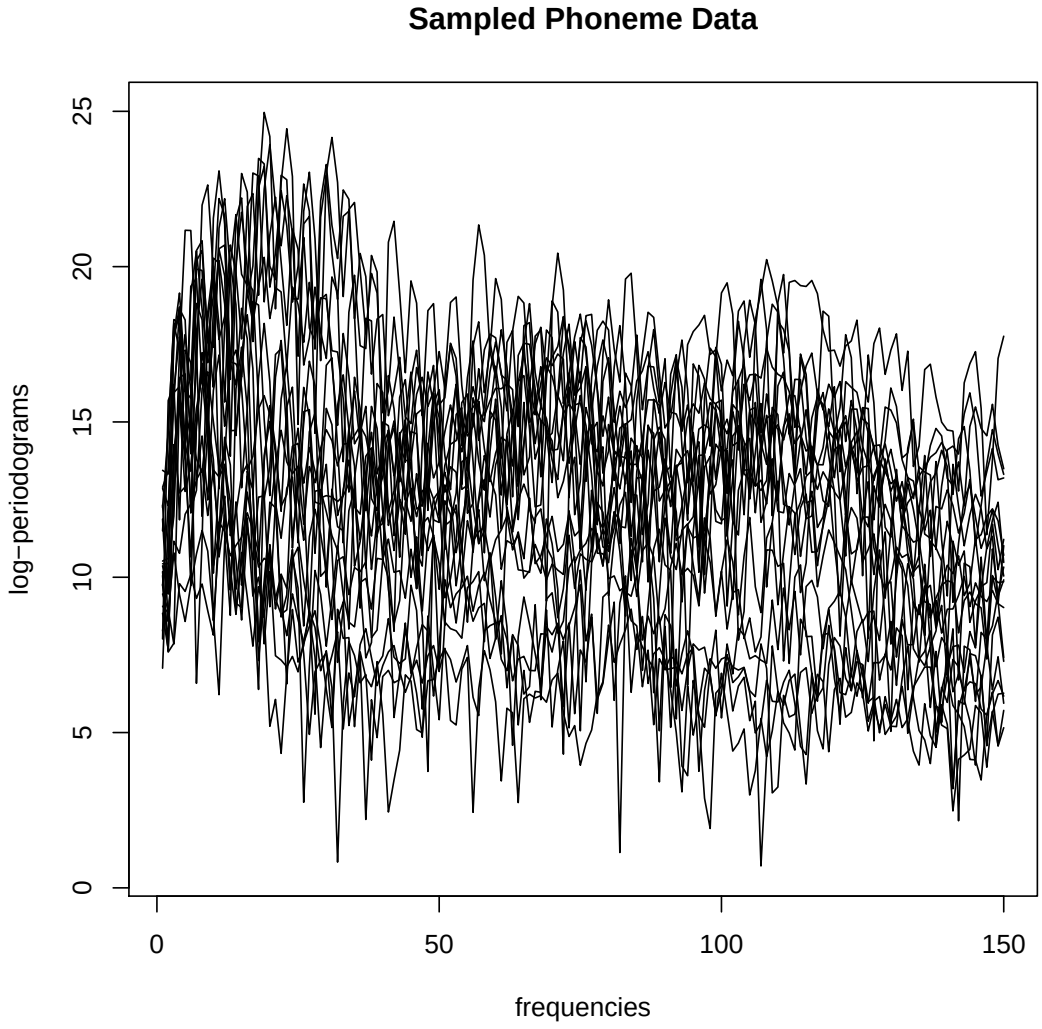


Figure 1.1: Sampled Phoneme data set as a function of wavelengths.

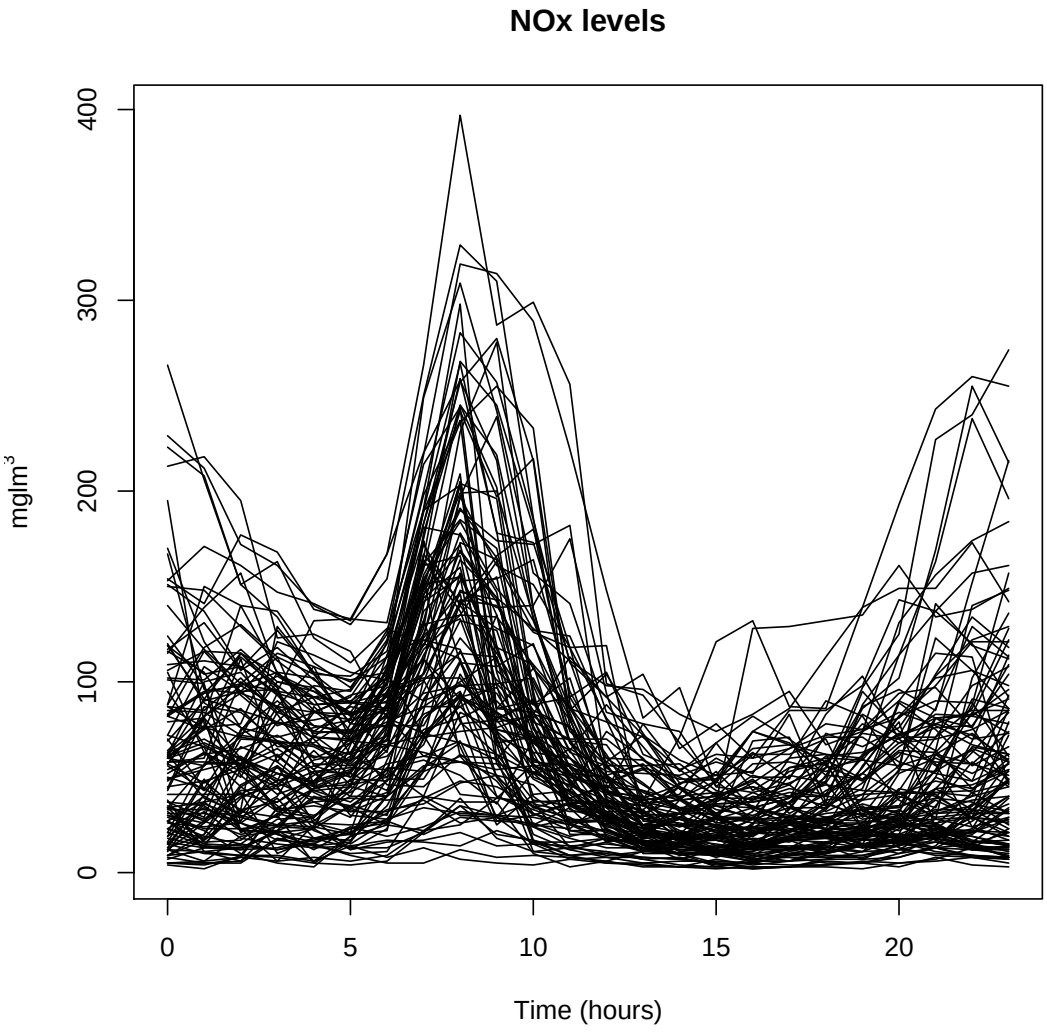


Figure 1.2: NOx levels measured every hour by a control station in Poblenou in Barcelona.

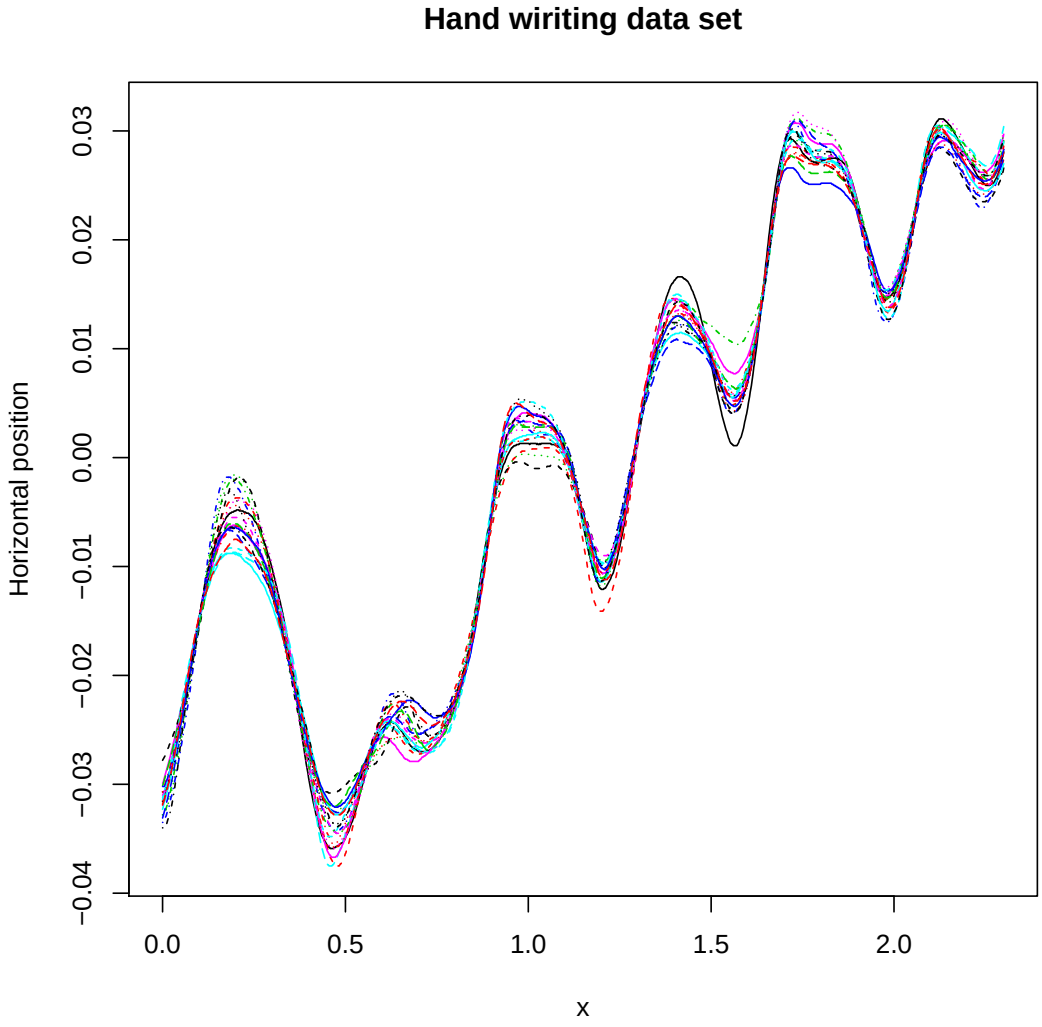
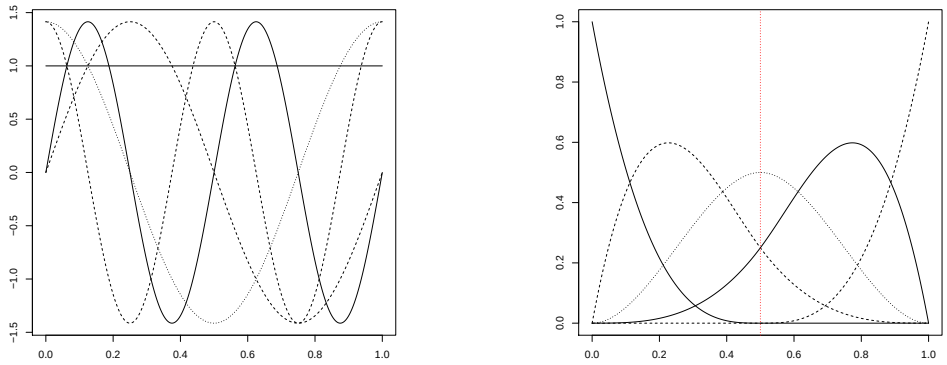


Figure 1.3: Handwriting data set x coordinates.



(a) (b)

Figure 1.4: Examples of Fourier (a) and B-spline (b) basis with number of basis = 5. The time interval for observation is $[0, 1]$

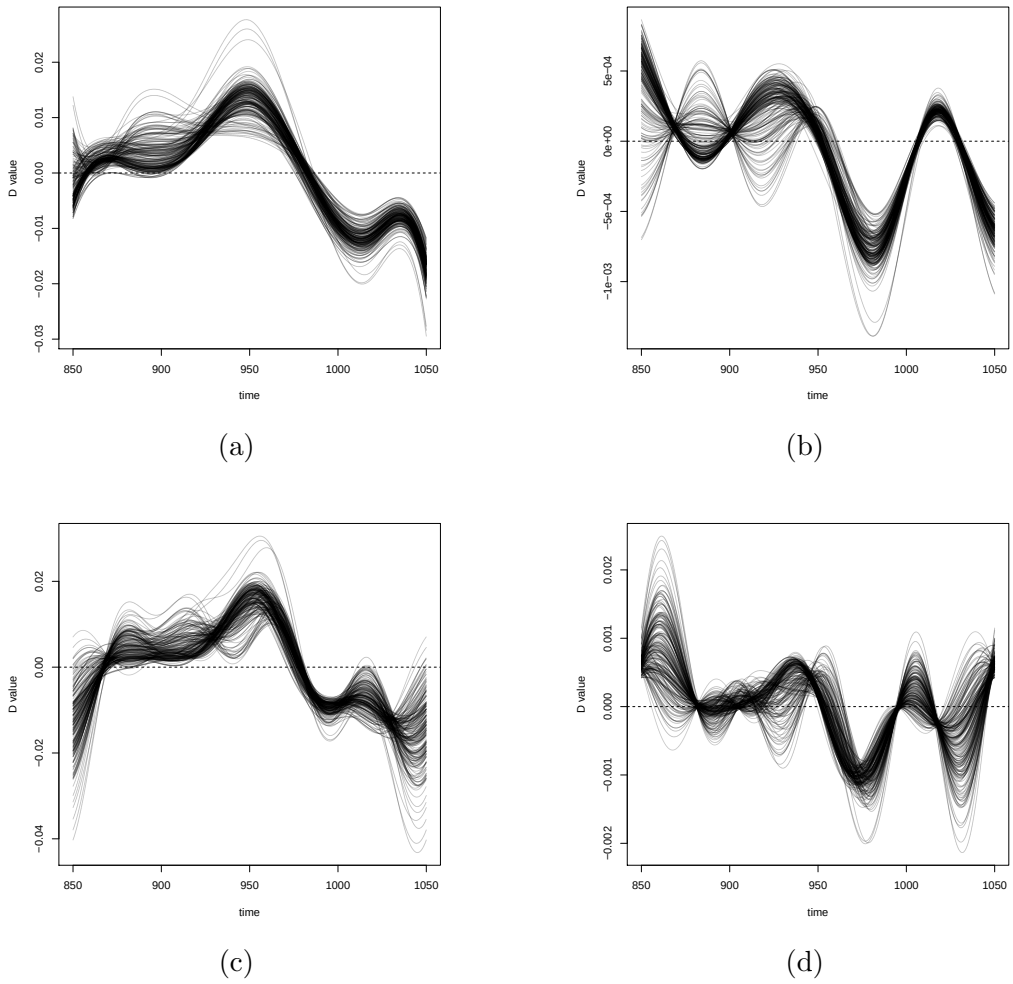


Figure 1.5: Derivative distance for Tecator data with (a) B spline basis order 1, (b) derivatives of order 2, (c) Fourier basis with derivative order 1, (d) order 2.

Chapter 2

Hierarchical Bayesian Modeling for Functional Data

2.1 Introduction

In this section, we are considering a Bayesian analysis of a functional data classification problem. We will focus on applications in medical imaging, specifically, experiments from motion sensor data and another experiment with EEG. These data sets can be considered as functional in nature as they all have one or more independently varying parameter. However, because of the nature of the experiments, these data sets would also have more complicated generative process, which would require careful modeling. In describing the data sets we will discuss the possible approaches we have considered in detail.

Bayesian methods provide a flexible way of approaching statistical problems, that makes it easy to deal with uncertainty in various levels. As opposed to frequentist methods, where there is no general way of leveraging the prior knowledge about the parameters, Bayesian methods are more flexible. Even though theoretically they have been very well developed for long time, they were relatively less used because of computational burden. With the improvement in computing technology, Bayesian methods are being used more and more in recent times. In statistics, one of the earlier works that triggered the popularity of Bayesian methods is by Gelfand and Smith [12], which focused on computation for Gibbs sampling. Several other papers describing computations and properties of MCMC algorithms were also very influential. Several works by Albert and Chib [1] and Chib [7] and several other authors considered implementation strategies and properties of Gibbs samplers.

There has also been several major works on understanding properties of Markov

chain Monte Carlo methods. Methods established by Chan and Geyer [6] for generating inference from Markov chains has been widely used. Techniques for studying convergences for chains under various conditions arising from statistical models has been explored in detail by Roberts and Tweedie [23] and Mengersen and Tweedie [19]. Works by Jarner and Tweedie [15] establishes conditions for verifying central limit theorem for certain Markov chains.

Algorithmic advances, studies of new techniques on efficient implementation of Markov chains both in terms of computation and storage efficiency has also been studied more recently. The idea of coupling has been used in several works and the theoretical properties of coupling were explored in Roberts and Rosenthal [24]. Using regenerative sampling to possibly use multiple chains to speed up computation has been tried out by Mykland et al. [21], however there are still no general recipe of using regenerative sampling for Markov chains. Some algorithmic tricks like optimized storage and reusing computations like prefetching has been studied by Brockwell [4]. For the case where the problem can be decomposed into multiplicative elements, it is possible to optimize the space use by splitting the data set into different machines. Similar schemes has been considered by Scott et al. [26] and Wilkinson [28].

2.2 Model

In this section, we will consider the model for classification of functional data. We are considering a hierarchical Bayesian model here. First, we briefly go over the

structure of the data.

2.2.1 Data Description

We consider functional data that is collected from medical imaging studies. The observations of these data sets would often be responses of some physical quantity with time. The data would most often have a spatial component associated with it, as there would typically by multiple sensors for the study. We will use the term multiple channel to denote the fact that data is being generated from multiple sources. Besides being a functional data set, these data would also present several general challenges. For example, it is common to have some correlation among the multiple channels. In many cases, these correlations are very important to explain. For example, in the case of an EEG data set, the channels would be multiple sensors planted on outside of the skull of the patient. If we are studying the electrical response of the Brain of a physical stimulation, then it is important to understand the dynamics of the activation of the brain regions, indicated by the channels. Besides the issue with channels, in most experiments there will be multiple subjects and multiple observations for each subject. It is natural to assume that the responses form one person would be more similar than responses from some one else. To account for the subject wise effect, we would use the hierarchical model, as we describe further.

2.2.2 A simple generative model

We perform a simple experiment to study the error rates of our estimates for our model. We generate data from the following hierarchy.

$$(X_{ij}|U_i, \Sigma_i) \sim N_p(U_i, \Sigma_{ij}) \tag{2.1}$$

$$(U_i) \sim N_p(\mu, \Gamma) \tag{2.2}$$

Here, we are assuming there are I individuals and for ith individual, we are obtaining J_i number of samples. The individual effect is modeled using the individual level intercept term U_i. The prior distribution of U is assumed to be normal. The conditional distribution of the data given a specific individual is also given by a normal distribution.

As this model hierarchy is quite common, and we are in the conjugate prior family of distributions, we can directly write the posterior distribution of the parameters. We will estimate the posterior means using the estimators given below. Firstly, the distribution of the sample means are written as

$$\bar{X}_i = \frac{1}{J_i}\sum_{i}^{n} X_{ij} \tag{2.3}$$

$$(\bar{X}_i) \sim N_p(\mu, J_i^{-1}\Sigma_i + \Gamma) \tag{2.4}$$

In this experiment, we estimate the parameter μ using the empirical version of the posterior estimates of the intercepts.

$$\hat{\mu} = \frac{1}{n}\sum_{i}^{n} \bar{X}_i \tag{2.5}$$

We then compare the results for changes in number of individuals I, with number of observations per individual and length of observation p. Here for simplicity, we have assumed that same number of observations were generated from each individual. In figure 2.1 we show the comparison of errors in estimating posterior mean with changes in dimension of the problem and with varying number of individual and with changing number of observations from each individual. It is clear that as the dimensionality of the problem increases, the average error increases, which is not surprising. With increase in number of individuals also, we see average error is increasing. This is somewhat counter intuitive, but as this result is averaged over all other parameters, this can be expected. The error against the number of observations does not seem to change much with increase of observations.

In figure 2.2, we compare the results for estimating the intra subject co-variance matrix for changes in number of individuals, number of observations and length of each observation. We are displaying the average error in Frobenius norm for the estimated co-variance matrix. The changes with dimension is very prominent as the error keeps increasing with dimension. For number of individuals and number of observations, the error does not change after certain point.

The comparison of errors for estimating the inter subject covariance matrix is shown in figure 2.3. The story about changes in various scenarios remains almost similar to other quantities in this experiment. Especially, the changes with the dimension is close to the pattern from the changes in errors for mean estimation in figure 2.1.

2.2.3 Effect of Basis Choice

Similar to the previous experiment, we have used the hierarchical model to generate the coefficients that we have used to generate functional data sets with. We have used two basis families for this experiment. The Fourier and B spline basis are both widely used families in the literature of functional data analysis and non parametric statistics. So, we have chosen to use those two families. The process of generation of the functional data from the given coefficients is simply performing the product of the basis components with the corresponding coefficients. As we are using the same set of coefficients with two different basis functions for this experiment, we will get completely different functional data sets. So, while recovering these coefficients in the estimation process, it is important to correctly specify the basis family. Later we will also study the effect of mis-specifying the basis family. Off course, in real data analysis the true basis family is only assumed and we will be studying the effect of multiple basis on the accuracy of our estimation or prediction process.

In this experiment, we simulate the coefficients as the model hierarchy and use two basis families to generate functional data sets. Then in our estimation process, we use the data sets as observed functional data and try to estimate the original parameters in the generation of the coefficients. As with the previous experiment, we change the number of coefficients, number of individuals and number of repetitions in a grid and report the average error in estimation of the parameters. The recovery of the coefficients from the functional data is almost perfect, and as we have used the pre specified basis for estimation, we only report

the results from one basis family. In this case in figure 2.4 we are reporting the results from Fourier basis, the story of the Bspline basis is very similar.

We show the two sets of functional data for one combination of the parameters for illustration in figure 2.5. These two sets of functional data is created from the same set of coefficients generated with the model, while they have used Fourier and Bspline basis family respectively. As we can see these two data sets look completely different, so in the modeling of real data analysis, it will be important to select the family and the number of coefficients that achieve the best result.

In figure 2.10, we show the comparison of the errors in estimating the posterior variance among individuals from the observed functional data simulated using the Fourier basis family. The comparisons are with changing number of coefficients, number of individuals and changing number of observations per individuals. For this study, we have chosen the number of coefficients between 5 to 50, with number of individuals varying from 10 to 200 and number of observations per individuals also between 10 to 200. The error accumulates with increase in number of coefficients. With increasing number of individuals, the error becomes stable after about 50 individuals, which is expected as the dimension of the coefficients is capped at 50. Similar thing happens with the number of observations as well.

In figure 2.6, we show the comparison of the errors in estimating the posterior covariance matrix among repeated observations from the same individuals. We are comparing the Frobenius norms between the true parameter and the estimated covariance matrix here. The effect of increasing coefficients seems to increase the error in estimation. This is expected as increasing coefficients also increases the

dimensionality of the problem. While in real data analysis, this will lead to a trade off between better modeling the data with more coefficients and reducing the number of coefficients to achieve minimum error. For the number of individuals, the error becomes stable after about 50 individuals. And for the number of repetitions of observations for each individual, the error does not increase and seems to be stable after 50 repetitions.

2.2.4 Brief Review of Precision Estimation

As in our estimation process, we are using several precision matrices, it is worthwhile to take a look at the problem of precision matrix estimation. This problem will also be relevant for several methods of classification, specifically discriminant analysis. While the problem of high dimensional precision matrix estimation is a very active research area, we will be dealing with the coefficients of the basis functions in most of the cases. So, the number of dimension will depend on the model chosen rather than the functional data set itself.

The problem of estimating the covariance matrix or the inverse covariance matrix is a classical problem in multivariate statistics. In it's basic form, the problem can be stated as estimation of the covariance matrix or it's inverse from a multivariate iid random vectors. As these are iid samples, the sample covariance matrix or the inverse of it is always a choice. Under the multivariate normal model, sample covariance matrix is also the maximum likelihood estimator. The properties of the sample covariance matrix is well understood and has been explored in Anderson [2] and Muirhead [20] in detail.

While we can use the sample covariance matrix based estimator, in several cases, the number of coefficients that we choose to fit in our functional data sets, exceeds the number of observations. So, in these cases we shift into the the domain of high dimensional covariance and precision matrix estimation. In these cases, the number of estimable parameters is large, as they are directly linked with the dimension of the problem. The typical way of dealing with this problem is by assuming some structural restriction in the data. Several methods propose assuming sparsity in the elements of the inverse covariance matrix.

The assumption of sparsity in the precision matrix also gives rise to interesting interpretation of the model. As, we model the data $X \sim N_p(\mu, \Sigma)$, then the precision matrix $\Omega = \Sigma^{-1}$ represents the conditional dependence among the coordinates $\{X_1, \cdots, X_p\}$. So, a zero in an element of the precision matrix would imply that coordinate is conditionally independent of other coordinates. Thus $\Sigma^{-1}_{ij} = 0$ would indicate $X_i \perp X_j | X_{\setminus\{i,j\}}$. This property of multivariate normal distribution has been exploited in various fields of application including genetics, medical imaging, and econometric modeling.

A rich literature exists on Gaussian graphical models, in which various type of sparsity inducing assumption is made while estimating the precision matrix. The original paper proposing the conditional independence property is by Dempster [8]. In high dimensional settings, Meinshausen and Buhlmann [18] have proposed reformulating the problem as system of sparse linear regression problem by conditioning on the coordinates. They have proposed solving the set of sparse regression problems using Lasso. The method by Banerjee and Ghaoui [3] proposed direct

estimation of the precision matrix using a penalized likelihood approach. The method by Yuan and Lin [31] also uses a model selection based approach. Graphical Lasso by Friedman et al. [11] also uses likelihood method with a penalty similar to lasso in regression. Some later methods like graphical Dantzig selector by Yuan [32] and CLIME by Cai et al. [5] proposed the use of linear programming for estimating the precision matrix. The theoretical properties of the estimators have been studied by Rothman et al. [25] and Ravikumar et al. [22].

2.3 Problem of Hierarchical Classification from Multichannel Data

Here we present the problem of Hierarchical classification form Multichannel functional data. We will see instances of such problems in our case study on activity recognition problem. Many other medical diagnostic problems can also be framed in a similar way. The *hierarchical* part of the problem comes in because of the effect of subject in the classification. In usual problem of classification, the dataset does not have any other identification than just belonging to a particular class. But, if the different data points originate from smaller number of sources but with multiple instances, then we might expect an effect of the source on the data, and that should be accounted for in the classification problem.

We collect the data from I source, in our case, they are different individuals. Each source produces multiple instances of data, ith source can have J_i number of instances. Each instance has K parallel streams of data. We consider labeled

data, so let Y_{ij} denote the label for observation from individual i and instance j. $Y_{ij} \in \{1, \cdots, L\}$. $i \in \{1, \cdots, I\}$, $j \in \{1, \cdots, J_i\}$, $k = \{1, \cdots, K\}$. So, our full dataset can be represented as : $(\mathcal{X}_{ijk}(t), y_{ij}); t \in (0, 1))$, where $\mathcal{X}(.) \in \mathcal{L}_2[0, 1]$.

In our dataset, the response is indicated by $(\mathcal{X}_{ijk}(t), y_{ij}); t \in (0, 1))$, where $\mathcal{X}(.) \in \mathcal{L}_2[0, 1]$. The data collected over a very fine grid of time points to ensure the functional nature of the data. In terms of the basis representation mentioned above, we can decompose each functional observation for any class as

$$(\mathcal{X}_{ijk}(t)|y_{ij} = l) = \sum_{a=1}^{A} W_{ijka} B_a(t) + R_{ijkl}(t)$$

Where, $\{B_a(.), a = 1, 2, \cdots\}$ is the family of basis functions, which has been decided to be used. $R_{ijkl}(.)$ denote the reminder term, which we choose to ignore and use only the first term in the above decomposition as a representation of the functional observation. Once the family of basis functions has been specified, we only need to keep track of the coefficients W_{ijka}. We have effectively reduced our data to a space of vector of coefficients which is $P = AK$ dimensional. We vectorize over the channels as will be explained below.

2.3.1 Estimation

We consider a Hierarchical model starting from each sources.

$$\text{Define } P = AK$$

$$(Z_{ij}|Y_{ij} = l) = Vec((W_{ijka}|Y_i j = l)_{k=1,\cdots,K;a=1,\cdots,A})$$

We can now model our data or the vectorized coefficients in the form of a hierarchy according to the way they have been collected. Each vector is assumed to be iid normal with parameters dependent on the individual. The means and variances for individual i is U_i, Σ_i. This would model the effect of the individual.Each source mean is U_i with a Gaussian prior with mean μ and covariance Γ.

$$(Z_{ij}|U_i, \Sigma_i, Y_{ij} = l) \sim_{iid} N_P((U_i|Y_{ij} = l), (\Sigma_i|Y_{ij} = l));$$
$$i = (1, \cdots, I)$$
$$(U_i|Y_{ij} = l) \sim_{iid} N_P(\mu|Y_{ij} = l, \Gamma|Y_{ij} = l)$$

2.3.2 Classification

With the above setup, we can find the estimates of the posterior means and variance of $(U_i|Y_{ij} = l)$ as, for a given class label,

$$\bar{Z}_i = J_i^{-1} \sum_j Z_{ij}$$
$$(\bar{Z}_i|U_i, \Sigma_i) \sim N_P(U_i, J_i^{-1}\Sigma_i)$$
$$(\bar{Z}_i) \sim N_P(\mu, J_i^{-1}\Sigma_i + \Gamma)$$

The posterior can be written as:

$$(U_i|\bar{Z}_i, \Sigma_i, \mu, \Gamma) \sim N_P(\theta_i, V_i)$$
$$V_i^{-1} = (J_i\Sigma_i^{-1} + \Gamma^{-1})$$
$$\theta_i = V_i(J_i\Sigma_i^{-1}\bar{Z}_i + \Gamma^{-1}\mu)$$

Means and variance estimations are done in the following way:

$$\begin{aligned}
\hat{\mu} &= I^{-1} \sum_i \bar{Z}_i \\
\hat{\Sigma}_i &= \sum J_i^{-1} (Z_{ij} - \bar{Z}_i)(Z_{ij} - \bar{Z}_i)^T \\
J_i^{-1}\hat{\Sigma_i} + \Gamma &= I^{-1} \sum (\bar{Z}_i - \hat{\mu})(\bar{Z}_i - \hat{\mu})^T \\
\hat{\Gamma} &= I^{-1} \sum [J_i^{-1}\hat{\Sigma_i} + \Gamma - J_i^{-1}\hat{\Sigma}_i]
\end{aligned}$$

All of these estimates are used to construct the posterior means and covariances θ_i and V_i.

Combining everything above, we produce the algorithm to classify new observations into one of the classes as follows:

1. **Dimension Reduction**: Fit a basis function and extract coefficients.
2. **Estimation**: Estimate the posteriors from the Bayesian Hierarchical Model.
3. **Rank based Estimation**: Use rank estimates of covariances and apply the NPN transformations.
4. **Regularized Discriminant Analysis**: Plug in the means and covariances into RDA discriminant function.
5. **Crossvalidation** Cross-validate for two tuning parameters of RDA, or use pre-determined values.

2.3.3 Properties

We will consider the sub problems of estimating the posterior mean for the hierarchical model. Given an hierarchical model of p variate normal distribution,

$$(Z|U, \Sigma) \sim N_p(U, \Sigma) \tag{2.6}$$

$$(U|\mu, \Gamma) \sim N_p(\mu, \Gamma) \tag{2.7}$$

We write the posterior distribution of the global mean as follows

$$(U|Z, \mu, \Gamma, \Sigma)\ N(\theta, V) \tag{2.8}$$

It is simple the derive the expression of θ and V, as we are in a conjugate family. However, for the sake of completion we present their expression.

Lemma 3. *Under the model in equation* (2.6)*, the estimates of the posterior parameters are given by*

$$V^{-1} = (\Gamma^{-1} + J\Sigma^{-1}) \tag{2.9}$$

$$\theta = V^{-1}(\Gamma^{-1}\mu + J\Sigma^{-1}\bar{Z}) \tag{2.10}$$

Proof. We can simply write the posterior of U with known Γ and μ. This will be

proportional to the joint distribution.

$$
\begin{aligned}
& f(U|Z,\mu,\Gamma,\Sigma) && (2.11)\\
\propto\ & \exp\big(-\frac{1}{2}(J(U-Z)^T\Sigma^{-1}(U-Z)+(U-\mu)^T\Gamma^{-1}(U-\mu))\big) && (2.12)\\
\propto\ & \exp\big(-\frac{1}{2}(U^T(J\Sigma^{-1}+\Gamma^{-1})U)-U^T(J\Sigma^{-1}Z+\Gamma^{-1}+\Gamma^{-1}\mu) && (2.13)\\
& -(Z^TJ\Sigma^{-1}+\mu^T\Gamma^{-1})U\big) && (2.14)\\
\propto\ & \exp\big(-\frac{1}{2}(U^T(J\Sigma^{-1}+\Gamma^{-1})U)-2U^T(J\Sigma^{-1}Z+\Gamma^{-1}+\Gamma^{-1}\mu)\big) && (2.15)\\
\propto\ & \exp\big(-\frac{1}{2}(U-\theta)^TV^{-1}(U-\theta)\big) && (2.16)
\end{aligned}
$$

Where, $\theta = V^{-1}(\Gamma^{-1}\mu + J\Sigma^{-1}\bar{Z})$ and $V^{-1} = (\Gamma^{-1} + J\Sigma^{-1})$. The fourth line comes because the second term in above in line 3 is same as it's transpose. This gives the posterior distribution. □

As we have the posterior estimate of the global mean, we now try to estimate the individual specific means. We propose a biased estimate that shrinks the estimate towards the prior mean, similar to James Stein estimator. One straight forward estimator can also be the mean of observations for each individual.

$$
\begin{aligned}
\tilde{U} &= \bar{Z} && (2.17)\\
\hat{U} &= [1-(p-2)\hat{\Sigma}^{-1}]\bar{Z} && (2.18)
\end{aligned}
$$

To study the properties of these estimators, we will use the squared error loss to calculate their risk. We will follow the method introduced by Stein [27] and James and Stein [14].

Theorem 4. *Under the case when global mean is* 0 *and global covariance matrix* I_p*, the empirical Bayes estimator dominates the maximum likelihood based estimator in terms of risk function under squared error loss.*

$$E_U||\hat{U} - U||^2 \leq E_U||\tilde{U} - U||^2 \tag{2.19}$$

where, the expectation is taken with respect to the distribution of U*.*

Proof. Let, $\hat{U}$ be an estimator of U that is absolutely continuous function of Z. The following identity holds for each component Z_α of $Z \in \mathbb{R}^p$. For $\alpha \in \{1, \cdots, p\}$

$$\begin{aligned}
(Z_\alpha - \hat{U}_\alpha)^2 &= [(Z_\alpha - U) - (\hat{U}_\alpha - U)]^2 && (2.20)\\
&= [(Z_\alpha - U_\alpha)^2 + (\hat{U}_\alpha - U)^2 - 2(Z_\alpha - U_\alpha)(\hat{U}_\alpha - U)] && (2.21)\\
(\hat{U}_\alpha - U)^2 &= (Z_\alpha - \hat{U}_\alpha)^2 - (\hat{U}_\alpha - U)^2 + 2(Z_\alpha - U_\alpha)(\hat{U}_\alpha - U) && (2.22)
\end{aligned}$$

Taking sum of all the coordinates $\alpha = \{1, \cdots, p\}$, and taking expectations with respect to distribution of Z,

$$\begin{aligned}
||U - \hat{U}||^2 &= ||Z - \hat{U}||^2 - ||Z - U||^2 + 2\sum_{\alpha=1}^{p}(Z_\alpha - U_\alpha)(\hat{U}_\alpha - U_\alpha) && (2.23)\\
E_U||U - \hat{U}||^2 &= E||Z - \hat{U}||^2 - E||Z - U||^2 + 2cov(Z, \hat{U}) && (2.24)
\end{aligned}$$

We now consider the problem in one dimension at first for simplicity. Later, we will consider the general case. We then observe the covariance term of the

above identity can be decomposed as follows.

$$
\begin{aligned}
E_U U\hat{(Z)}' &= \int_{-\infty}^{\infty} \hat{U}'(Z)\phi(Z)dZ && (2.25)\\
&= \phi(Z)\int_{-\infty}^{\infty} \hat{U}'dZ - \int_{-\infty}^{\infty}(\phi'(Z)\int \hat{U}'(Z)dZ)dZ && (2.26)\\
&= [\phi(Z)\hat{U}(Z)]_{\infty}^{\infty} - \int_{-\infty}^{\infty}\phi'(Z)\hat{U}(Z)dZ && (2.27)\\
&= \int_{-\infty}^{\infty} Z\phi(Z)\hat{U}(Z)dZ && (2.28)\\
&= E[Z\hat{U}] && (2.29)\\
&= cov(Z,\hat{U}) && (2.30)
\end{aligned}
$$

Alternatively, we can also use a different decomposition for the same identity in one dimension. We will use the following result for decomposition of expectation

$$
\begin{aligned}
E_U U\hat{(Z)}' &= \int_0^{\infty} U\hat{(Z)}'\big(\int_Z^{\infty} Z\phi(Z)dZ\big)dZ \int_{-\infty}^{0} && (2.31)\\
&\quad -U\hat{(Z)}'\big(\int_{-\infty}^{Z} Z\phi(Z)dZ\big)dZ && (2.32)\\
&= \int_0^{\infty} Z\hat{U}(Z)\big(\int 0^{Z}\hat{U}'(Z)dZ\big)dZ && (2.33)\\
&\quad -\int_{-\infty}^{0} Z\hat{U}(Z)\big(\int Z^{0}\hat{U}'(Z)dZ\big)dZ && (2.34)\\
&= \int_0^{\infty} Z\phi(Z)[\hat{U}(Z)]_0^Z dZ - \int_{-\infty}^{0} Z\phi(Z)[\hat{U}(Z)]_Z^0 dZ && (2.35)\\
&= \int_{-\infty}^{\infty} Z\phi(Z)\hat{U}(Z)dZ && (2.36)\\
&= E[Z\hat{U}] && (2.37)\\
&= cov(Z,\hat{U}) && (2.38)
\end{aligned}
$$

For the multivariate version of the above identity, we consider the approach of

Stein **?**]. We will consider the terms of the matrix $E_U[\nabla \hat{U}(Z)]$. The estimator is considered coordinate wise as $\hat{U}(Z) = \hat{U}(Z_\alpha; Z_{-\alpha})$, where $Z_{-\alpha}$ indicate the vector $(Z_1, \cdots, Z_{\alpha-1}, Z_{\alpha+1}, \cdots, Z_p)$. So the function $\hat{U}(., Z_{-\alpha}) : \mathbb{R} \mapsto \mathbb{R}$. We can thus use the same method as in the univariate case.

$$
\begin{aligned}
E_U \frac{\partial \hat{U}}{\partial Z_\alpha}(Z_\alpha; Z_{-\alpha}) &= \int_\infty^\infty \frac{\partial \hat{U}}{\partial Z_\alpha}(Z_\alpha; Z_{-\alpha}) \phi(Z_\alpha) dZ_\alpha && (2.39)\\
&= \int_0^\infty \frac{\partial \hat{U}}{\partial Z_\alpha}(Z_\alpha; Z_{-\alpha}) \Big(\int_{Z_\alpha}^\infty t\phi(t)dt \Big) dZ_\alpha && (2.40)\\
&\quad - \int_{-\infty}^0 \frac{\partial \hat{U}}{\partial Z_\alpha}(Z_\alpha; Z_{-\alpha}) \Big(\int_{-\infty}^{Z_\alpha} t\phi(t)dt \Big) dZ_\alpha && (2.41)\\
&= \int_0^\infty Z_\alpha \phi(Z_\alpha) \Big(\int_0^{Z_\alpha} \frac{\partial \hat{U}}{\partial Z_\alpha}(Z_\alpha; Z_{-\alpha}) dZ_\alpha \Big) && (2.42)\\
&\quad \int_{-\infty}^0 Z_\alpha \phi(Z_\alpha) \Big(\int_{Z_\alpha}^0 \frac{\partial \hat{U}}{\partial Z_\alpha}(Z_\alpha; Z_{-\alpha}) dZ_\alpha \Big) && (2.43)\\
&= \int_0^\infty Z_\alpha \phi(Z_\alpha) [\hat{U}(.; Z_{-\alpha})]_0^{Z_\alpha} d_{Z_\alpha} - && (2.44)\\
&\quad \int_{-\infty}^0 Z_\alpha \phi(Z_\alpha) [\hat{U}(.; Z_{-\alpha})]_{Z_\alpha}^0 d_{Z_\alpha} && (2.45)\\
&= \int_{-\infty}^\infty Z \phi(Z) \hat{U}(Z) dZ && (2.46)\\
&= E[Z\hat{U}] && (2.47)\\
&= cov(Z, \hat{U}) && (2.48)
\end{aligned}
$$

We now use the above identity to estimate the risk of the second estimator $\tilde{U}$ in equation (2.19).

$$E_U||\tilde{U} - U||^2 = E||\bar{Z} - \tilde{U}||^2 - p/J + 2\sum_{\alpha=1}^{p} \frac{\partial \tilde{U}}{\partial Z_\alpha}(Z_\alpha; Z_{-\alpha})dZ_\alpha \quad (2.49)$$

$$= p/J \quad (2.50)$$

For the estimator $\bar{U}$ as proposed in equation (2.17), the risk can be written as

$$E_U||\hat{U} - U||^2 = E||\bar{Z} - \bar{U}||^2 - p/J + 2\sum_{\alpha=1}^{p} \frac{\partial \hat{U}}{\partial Z_\alpha}(Z_\alpha; Z_{-\alpha})dZ_\alpha \quad (2.51)$$

The first term is the expectation of the error from an unbiased estimator of U. As Σ_p is assumed to be identity, we can write this term as $(p-2)^2/||\bar{Z}||^2$. For the third term, we use the identity in equation (2.25) to reduce to $2(p/J - \sum_{\alpha=1}^{p}(p-2)^2/||\bar{Z_\alpha}||^2)$. Thus, the risk term in equation (2.51) is lower than that of equation (2.49).

□

2.3.4 Robust Estimation of Covariance

Discriminant analysis involves estimation of means and covariance matrices from the data. In fact, we will use estimates of both covariance matrix and precision matrix in the discriminant function. It is known that under the assumption of Gaussianity, the precision matrix can encode the dependency structure of the data, in the form of an undirected graph. In that case, for a random vector $(X_1, \cdots, X_p)^T$, the set of edges E denote the conditional dependency of each variable given everything else. In other words, if X_i is independent of X_j conditional to $X_{\setminus\{i,j\}}$, then $\Omega_{ij} = 0$, where 0 indicates absence of an edge.

Estimation of precision matrix and using that for building graphical models of the data is a well researched problem in statistics and computer science. In low dimension, inverse of the estimated covariance matrix is used as an estimate of precision matrix if it is not rank deficient. In high dimension, Meinshausen and Buhlmann [18] proposed a collection of regression problem as a solution of precision estimation. Penalized likelihood based methods developed by Yuan [30], Banerjee and Ghaoui [3], Friedman et al. [11] are also quite well known. Graphical Dantzig selector, CLIME Cai et al. [5] etc have also been proposed more recently. Many of these methods impose sparsity using various techniques.

To relax the assumption of Gaussian-ness of the data, Liu, Lafferty and Wasserman has proposed the Non-Paranormal family. A random vector $X = (X_1, \cdots, X_p)$ belongs to a Non-Paranormal family if there exists a set of univariate monotone functions $\{f_i\}_{i=1}^p$ such that $f(X) := (f_1(X_1), \cdots, f_p(X_p)$ is Gaussian. It is called $X \sim NPN_p(f, \Sigma)$ if $f(X) \sim N(0, \Sigma)$, where Σ denote the correlation matrix.

Because by definition, $f()$ is a set of univariate monotone transformations, rank correlation matrices of $f(X)$ and X are same. This idea has been used by Liu Lafferty and Wasserman to estimate Ω directly using rank based estimates of covariance matrix. There are two definitions of rank correlations in the literature, Spearman's ρ and Kendall's τ.

$$\hat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_i^j - \bar{r}^j)(r_i^k - \bar{r}^k)}{\sqrt{\sum_{i=1}^n (r_i^j - \bar{r}^j)^2 \sum_{i=1}^n (r_i^k - \bar{r}^k)}} \tag{2.52}$$

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le \alpha < \alpha' \le n} sign((x_j^\alpha - x_j^{\alpha'})(x_k^\alpha - x_k^{\alpha'})) \tag{2.53}$$

Using the results from Kruskal [16] we can connect the covariance in the NPN model with rank correlation coefficients. The following lemma is used in Liu et al. [17].

Lemma 5. *Assuming* $X \sim NPN(0, \Sigma, f)$, $\Sigma_{jk} = 2\sin(\frac{\pi}{6}\rho_{jk} = \sin(\frac{\pi}{2}\tau_{jk})$

Based on the above connection, Liu et al. [17] have proposed the following estimators of the correlation matrix:

$$\hat{S}^{\rho}_{jk} = \begin{cases} 2\sin(\frac{\pi}{6}\hat{\rho}_{jk}), & , j \neq k \\ 1, & , j = k \end{cases} \tag{2.54}$$

Similarly for Kendal's τ,

$$\hat{S}^{\tau}_{jk} = \begin{cases} \sin(\frac{\pi}{2}\hat{\tau}_{jk}), & , j \neq k \\ 1, & , j = k \end{cases} \tag{2.55}$$

Liu Lafferty and Wasserman have proposed using the above estimates $\hat{S}^{\rho}$ and $\hat{S}^{\tau}$ in the optimization problem for several regularized methods in the literature to estimate the precision matrix. They have studied the properties of the final estimator and both of the above estimators have similar properties. Their result states that the rate of convergence of the rank based precision estimate remains the same as the original estimator for several methods of precision estimation including parametric Graphical Lasso, Graphical Dantzig selector or CLIME. Their simulation results show the performance of the precision estimate under Gaussian setup and under several families of transformations and in the presence of outliers.

Based on the performance of the estimators, they recommend the use of these estimates even when the normality assumption is valid, as the estimates seem to perform quite competitively. In our method we use these rank based estimates. Although our method could use any of the above estimates, because of ease in computation we use the estimate based on Spearman's ρ in our method.

Handling possible rank deficiencies

While we are operating in the coefficient space, where coefficients attach to orthonormal basis family, the covariance matrix of the coefficients for each channel is unlikely to be rank deficient. However, as we are considering multiple channels, it is possible to have similar coefficients between channels, and resulting co-variance matrix could be rank deficient. We are estimating the posterior means and variances based on the estimated precision of the coefficients, hence we needed to address this issue. We consider the approach by **?**]. The regularized estimate of the covariance matrix is given by:

$$\hat{\Sigma}_g(\lambda, \gamma) = (1 - \gamma)\hat{\Sigma}_g(\lambda) + \frac{\gamma}{p} tr(\hat{\Sigma}_g(\lambda))I \tag{2.56}$$

Where (λ, γ) are a set of regularization parameters. $0 \leq \lambda \leq$ controls the degree of shrinkage of the individual class covariance matrix towards the pooled covariance matrix. So, in effect this is pushing the discriminant function towards the discriminant function of LDA. With $\lambda = 1$, this gives LDA. Here γ is an additional regularization parameter that controls the degree of shrinkage towards a multiple of identity matrix. This works similar to ridge regression and in effect

reduces the larger eigenvalues and increases the smaller ones.

We propose using the rank based estimate of the sample covariance matrix in the above equation to get an estimate of the precision matrix.

$$\hat{S}_g^{\rho}(\lambda,\gamma) = (1-\gamma)\hat{S}_g^{\rho}(\lambda) + \tfrac{\gamma}{p} tr(\hat{S}_g^{\rho}(\lambda))I \tag{2.57}$$

$$\hat{S}_g^{\tau}(\lambda,\gamma) = (1-\gamma)\hat{S}_g^{\tau}(\lambda) + \tfrac{\gamma}{p} tr(\hat{S}_g^{\tau}(\lambda))I \tag{2.58}$$

Using the regularized estimate of the precision matrix from the above equation, we plug it in the discriminant function:

$$\delta_g^{\rho}(X) = (X-\mu_k)^T \hat{S}[_g^{\rho}(\lambda,\gamma)^{-1}(X-\mu_g) + \tag{2.59}$$

$$\ln|\hat{S}_g^{\rho}(\lambda,\gamma)| - 2\ln(\pi_g) \tag{2.60}$$

$$\delta_g^{\tau}(X) = (X-\mu_k)^T \hat{S}_g^{\tau}(\lambda,\gamma)^{-1}(X-\mu_g) + \tag{2.61}$$

$$\ln|\hat{S}_g^{\tau}(\lambda,\gamma)| - 2\ln(\pi_g) \tag{2.62}$$

$$\tag{2.63}$$

Then finally the classification is done by assigning the class label which has the highest value of the discriminant function for a particular observation.

Simulation Study

We generate two classes from same mean function but the errors have different distributions. The two classes has same length of 200 time points each and we generate 1000 replications of each. The mean functions used are constant and

$\sin((2\pi/T)t)$. Gaussian noise was added to each class with equal variance. In one class the error was independent of each other, in other case, it had a dependence. The error structure is visualized in 2.11. The classification method used here are discriminant analysis with three kind of plug in estimator proposed above. We compared the misclassification rates with the nearest neighbor classifier based on a variety of distance mentioned in Ferraty [10]. The neighborhood based methods work on identifying the difference in shape between two classes, however, because we used the same mean function, they did not perform well compared to the discriminant analysis.

2.3.5 Activity Recognition Problem

One of the common uses of functional data analysis is in sensor data. It is becoming increasingly popular these days to use sensors and record information from everyday objects and activity. Such kind of data is used for making intelligent decision about resource allocation, identifying the source and classifying them into categories etc. Activity recognition is one such area which is an emerging field of research. The problem of activity recognition has been studied in video and audio stream analysis. It can also be based on classifying sensory data collected from one or many accelerometers mounted on the subjects. The equipment needed for such experiment is relatively cheap and easily deployable in large scale systems. In this case study we examine the data collected from one such experiment.

Description of the Data

The original data for this experiment is collected by Cassale et al. The dataset is publicly available via UCI machine learning repository. The original dataset had data collected from accelerometer mounted on chest of 15 subjects performing 7 activities. The dataset can both be used for classification of activity and user identification based on their motion pattern. The sampling frequency of the accelerometer was 52Hz, producing 116101 data points of acceleration information from all the users. The acceleration is a vector recorded as (a_x, a_y, a_z) in each of three directions. The activities were also labeled from 0 to 7.

So, in the notation of our problem formulation, we have number of individuals $I = 15$, number of channels $K = 3$. We have sampled from each subject to produce small segments of the activity pattern. The length of each segment is 250 time points, which is chosen keeping both statistical and computational considerations in mind. The sampled curves for each activity represent a small time interval of observation for that subject. The segment length, although here chosen by us, may not always be in the control of the experimenter, as this could be predetermined by the sensor or the algorithm. We have sampled the curves based on a common starting range of the response. This is done to maintain the phase difference relatively small between each sample from the same individual. Our method can handle the case when the number of sampled instances per individual is greater than one, and it is generally advisable to have higher sample size for better accuracy of the method. However, in this dataset, we were limited by the original observations. The length of the observations for each activity were not

fixed, and the two activity we have chosen to use for our classification, (walking and using the stair) had different number of observations. So, our sample size were mostly limited by the smaller of the two classes, (using the stairs in this case.)

2.4 EEG Data

Electroencephalography or EEG is a widely used technique to study electro physiological response of the brain in a non-invasive way. The response is electrical voltage measured from the electrodes placed on the scalp. The voltage fluctuates due to change in ionic current through the neurons in the brain. The data collected from different electrodes are usually called channels of the data. The number of electrodes can vary by experiment, however, there are standard conventions that determine positions of the electrodes. The analysis of EEG data is a broad subject. The nature of analysis varies from spectral content analysis to time domain study of averaged responses. EEG analysis is used for diagnosis of various brain conditions including during epilepsy, sleep disorder, coma etc. EEG is also used to measure brain response due to sensory or audio-visual stimuli. Such studies are referred to as Event Related Potential.

2.4.1 Description of the data

The dataset considered in our analysis is available in public domain through `https://archive.ics.uci.edu/ml/datasets/EEG+Database`. The data is collected in Neurodynamics Laboratory, State University of New York Health Center as a part of large study to examine how EEG correlates with genetic predisposition to alcoholism. There are 122 subjects from two groups, alcoholics and control. Each subject was given either a single visual stimuli or two stimuli of either matching pictures or non-matching pictures. Pictures were randomly chosen from 1980 Snodgrass and Vanderwart picture set. Each subject had 64 electrodes on their brain, the positions of the electrodes were at standard sites (Standard Electrode Position Nomenclature).

We show the mean response over all the trials for each of the stimuli in two groups. The image represents 64X256 matrix of response, where colors denote the voltage.

2.4.2 Analysis by channel

We first set up our notation for the data. Let there be I individuals and we will denote them by subscript $i \in \{1, \cdots, I\}$. The replicates of the data would be denoted by $j \in \{1, \cdots, J\}$. Channels would be denoted by $k = \{1, \cdots, K\}$. There was three experimental conditions, based on the kind of stimulus used, they would be denoted by $m \in \{1, \cdots, M\}$. The independent index for the functional data in our case is time. If the total time of observation is T, then, we denote each observation as an index $t \in \{1, \cdots, T\}$. In our data, $I = 122$, replicates J can be

$40 - 50$, number of channels $K = 64$, time for observations $T = 256$. There were $M = 3$ stimulus, and two classes of subjects $l \in \{1, \cdots, L\}$, or $L = 2$ present. We denote the response as $(V_{ij}^m(t,k)|Y_i = l)$, where Y_i denote the ith subject.

One way of handling this data would be to consider this as a bivariate functional data, where the response is a function of both time and channels. We can scale the time and channel axis accordingly to ensure that $(t,k) \in [0,1]^2$ for ease of analysis. Now, we can use a bivariate smoother to simultaneously smooth over both axes. Two popular approaches for bivariate smoothing are bivariate P-splines as proposed by Xiao et al. [29] and Eilers and Marx [9]. Another approach is to use thin plate splines as suggested in Hancock and Hutchinson [13]. A computationally efficient method based on P-splines is given by Xiao et al. [29], which uses a smoother matrix for rows and columns separately. Once the coefficients are extracted from the bivariate smoother, we suggest using a hierarchical model similar to the setup of the activity dataset.

Another way of analyzing this dataset would be to deal with each channels individually and employing a multiple testing procedure to identify the channels which are most distinguishing. Then, we can vectorize the data over those channels and use a basis function to reduce the dimension. The benefit of this approach is that the important channels would have some physical interpretation which is relevant to the medical community.

In our analysis, we have used Fourier basis to smooth the response for each observation. We have removed the observations with static response, as quite a few observations have shown delayed response. EEG studies can have constant

responses for a section of the time window, however, we have decided to remove such observations, as our model is not suited for these kind of responses. We have also removed the observations that had extreme values in them. Most of the EEG voltage response is expected to be stationary time series, while there were a few with extremely large or small observations.

We used the hierarchical model to treat the stimulus variable as the first level of hierarchy. There were three levels of stimulus in this experiment. So, for each levels of stimulus, there were multiple trials, each generating a full observation of EEG response. So, in this case the global means would indicate the mean of all three stimulus cases.

To generate the response for each channel by the group, we have averaged over all the individuals in each channel and shown the responses in the figure 2.18.

We have used similar method to summarize over all the individuals and trials to generate the mean response corresponding to each kind of stimulus for all the channels. The figure 2.19 shows the channel responses for three stimulus.

Bibliography

[1] James H. Albert and Siddhartha Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669, June 1993. ISSN 01621459. doi: 10.2307/2290350. URL `http://www.jstor.org/stable/2290350?origin=crossref`.

[2] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley

series in probability and statistics. Wiley-Interscience, Hoboken, N.J, 3rd ed edition, 2003. ISBN 978-0-471-36091-9.

[3] O. Banerjee and El Ghaoui. L. and d'Aspremont, A. (2008). *Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res.*, 9:485–516.

[4] A. E Brockwell. Parallel Markov chain Monte Carlo Simulation by Pre-Fetching. *Journal of Computational and Graphical Statistics*, 15 (1):246–261, March 2006. ISSN 1061-8600, 1537-2715. doi: 10.1198/106186006X100579. URL `http://www.tandfonline.com/doi/abs/10.1198/106186006X100579`.

[5] T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106:594–607, 2011.

[6] K. S. Chan and C. J. Geyer. Comment on âĂIJMarkov chains for exploring posterior distributions.âĂİ. *Annals of Statistics*, 22:1747–1758, 1994. bibtex: chan94.

[7] Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995. URL `http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476635`.

[8] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[9] Paul H. C Eilers and Brian D Marx. Generalized Linear Additive Smooth Structures. *Journal of Computational and Graphical Statistics*, 11(4):758–783, December 2002. ISSN 1061-8600, 1537-2715. doi: 10.1198/106186002844. URL http://www.tandfonline.com/doi/abs/10.1198/106186002844.

[10] Vieu Philippe Ferraty, FrÃl'dÃl'ric. *Nonparametric Functional Data Analysis*. 2006.

[11] J. H. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

[12] Alan E. Gelfand and Adrian F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):pp. 398–409, 1990. ISSN 01621459. URL http://www.jstor.org/stable/2289776. bibtex: gelfand1990 bibtex[jstor_articletype=research-article;jstor_formatteddate=Jun., 1990].

[13] P.A. Hancock and M.F. Hutchinson. Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. *Environmental Modelling & Software*, 21(12):1684–1694, December 2006. ISSN 13648152. doi: 10.1016/j.envsoft.2005.08.005. URL http://linkinghub.elsevier.com/retrieve/pii/S1364815205001659.

[14] W. James and Charles Stein. Estimation with Quadratic Loss. The Regents of the University of California, 1961. URL http://projecteuclid.org/euclid.bsmsp/1200512173.

[15] SÃÿren F. Jarner and Richard L. Tweedie. Necessary conditions for geometric and polynomial ergodicity of random-walk-type. *Bernoulli*, 9(4):559–578, 2003. doi: 10.3150/bj/1066223269. URL http://dx.doi.org/10.3150/bj/1066223269. bibtex: jarner2003.

[16] William H. Kruskal. Ordinal Measures of Association. *Journal of the American Statistical Association*, 53(284):pp. 814–861, 1958. ISSN 01621459. URL http://www.jstor.org/stable/2281954.

[17] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012. doi: 10.1214/12-AOS1037. URL http://dx.doi.org/10.1214/12-AOS1037.

[18] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34:1436–1462, 2006.

[19] K.L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24:101–121, 1996. bibtex: meng:twee:1996.

[20] Robb J. Muirhead, editor. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, March 1982. ISBN 978-0-470-31655-9 978-0-471-09442-5. URL http://doi.wiley.com/10.1002/9780470316559. DOI: 10.1002/9780470316559.

[21] Per Mykland, Luke Tierney, and Bin Yu. Regeneration in Markov Chain Samplers. *Journal of the American Statistical Association*, 90(429):233, March 1995. ISSN 01621459. doi: 10.2307/2291148. URL `http://www.jstor.org/stable/2291148?origin=crossref`.

[22] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. *Model selection in Gaussian graphical models: High-dimensional consistency of l1-regularized MLE*. In Advances in Neural Information Processing Systems 22. MIT Press, Cambridge, MA, 2009.

[23] G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996. bibtex: rt96.

[24] Gareth O. Roberts and Jeffrey S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475, 2007. doi: 10.1239/jap/1183667414. URL `http://dx.doi.org/10.1239/jap/1183667414`. bibtex: roberts2007.

[25] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008.

[26] Steven L. Scott, Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayes and big data:

The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016. URL http://www.tandfonline.com/doi/abs/10.1080/17509653.2016.1142191.

[27] Charles Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. The Regents of the University of California, 1956. URL http://projecteuclid.org/euclid.bsmsp/1200501656.

[28] Darren J. Wilkinson. Parallel bayesian computation. *Statistics Textbooks and Monographs*, 184:477, 2006. URL http://www.mas.ncl.ac.uk/~ndjw1/docs/pbc.pdf.

[29] Luo Xiao, Yingxing Li, and David Ruppert. Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):577–599, 2013. ISSN 1467-9868. doi: 10.1111/rssb.12007. URL http://dx.doi.org/10.1111/rssb.12007.

[30] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286, 2010.

[31] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.

[32] Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010. URL http://www.jmlr.org/papers/v11/yuan10b.html.

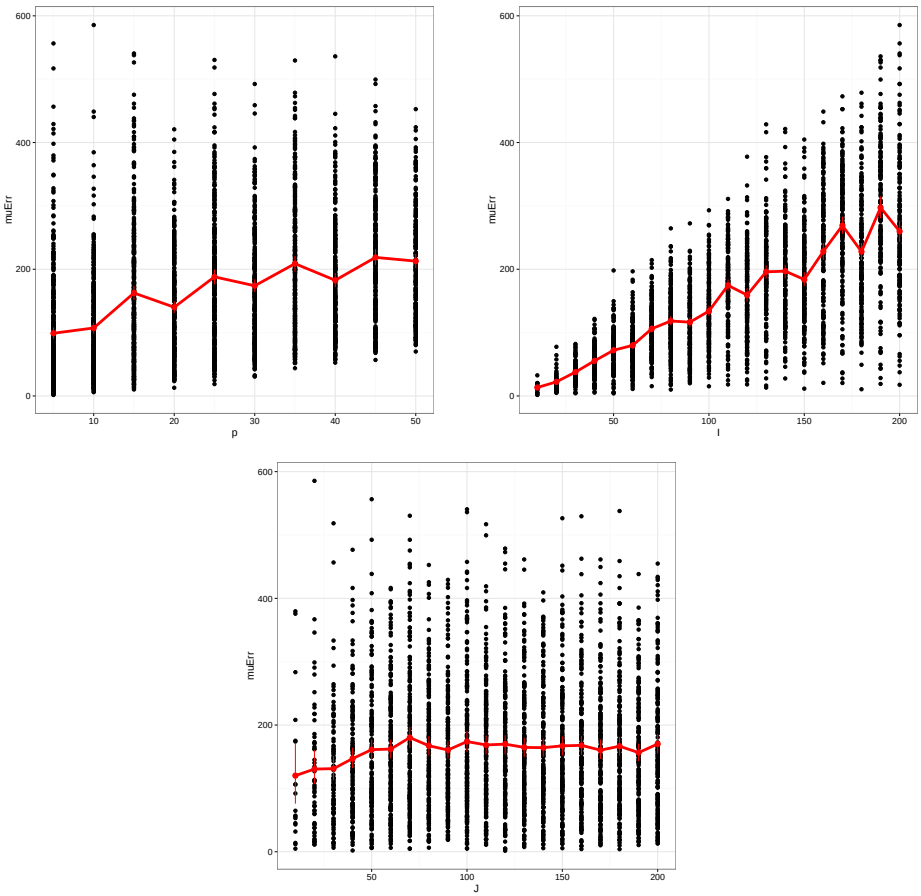


Figure 2.1: Comparison of the posterior estimate of mean with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

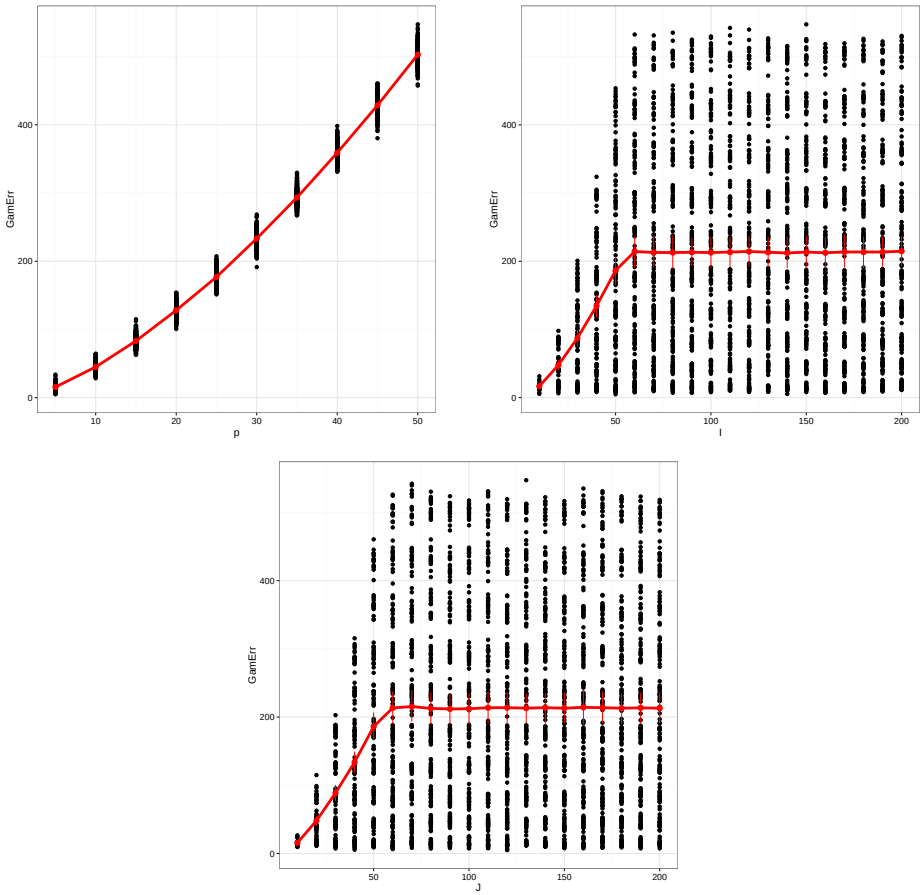


Figure 2.2: Comparison of the posterior estimate of Γ with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

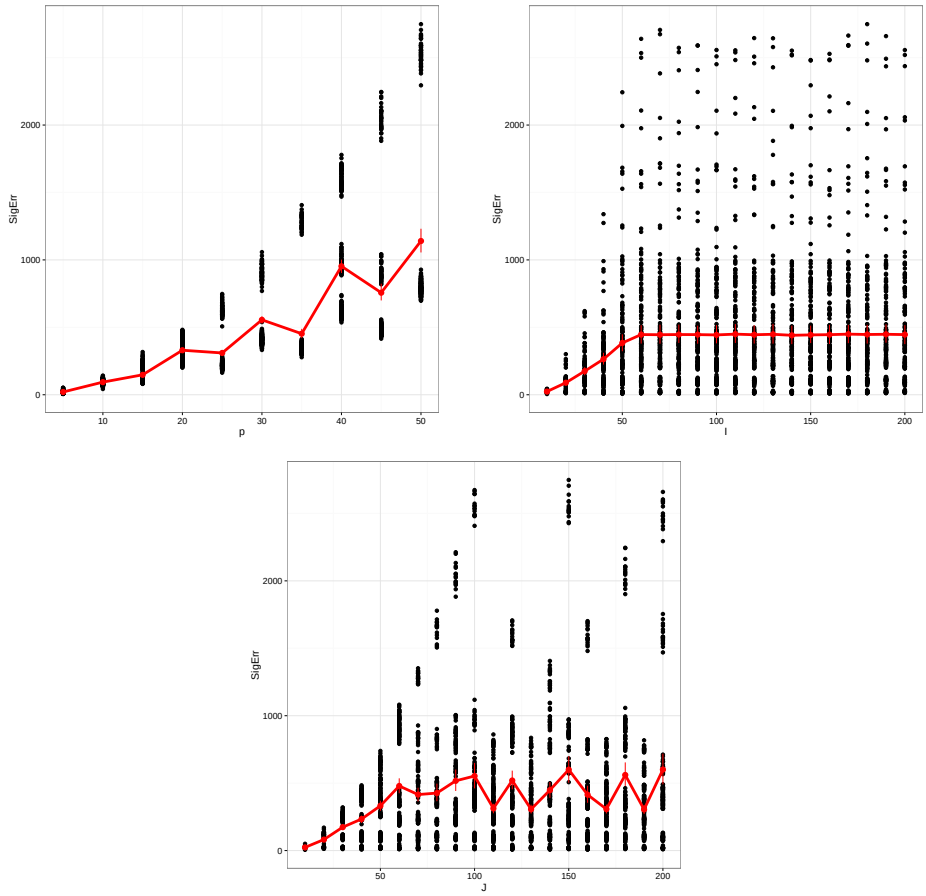


Figure 2.3: Comparison of the posterior estimate of inter subject co-variance matrix with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

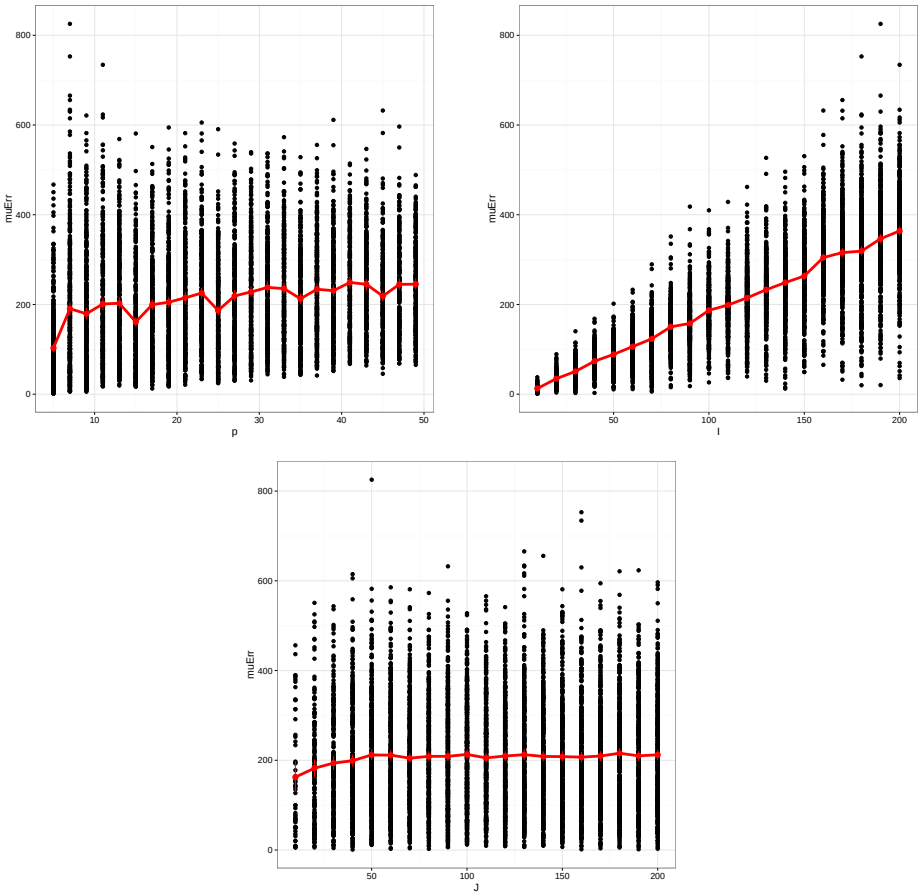


Figure 2.4: Comparison of the estimate of posterior means from functional data with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

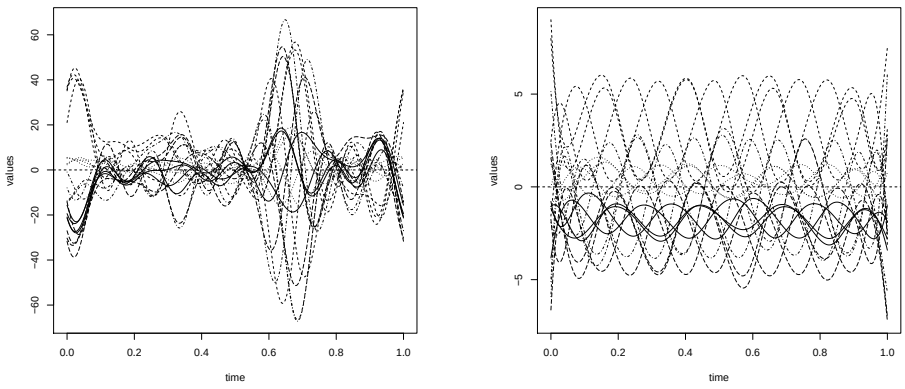


Figure 2.5: Comparison of the generated functional data for the same set of coefficients using (a) Fourier basis and (b) using the Bspline basis family.

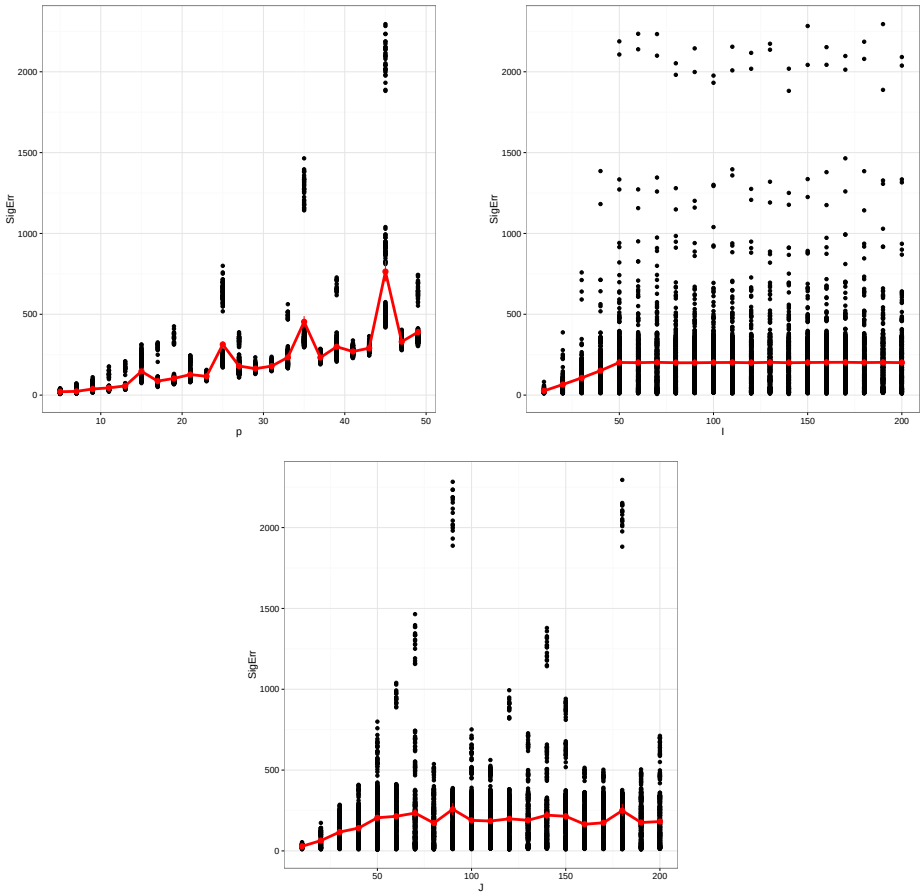


Figure 2.6: Comparison of the estimate of posterior variance among repeated observations from functional data generated from Fourier basis family with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

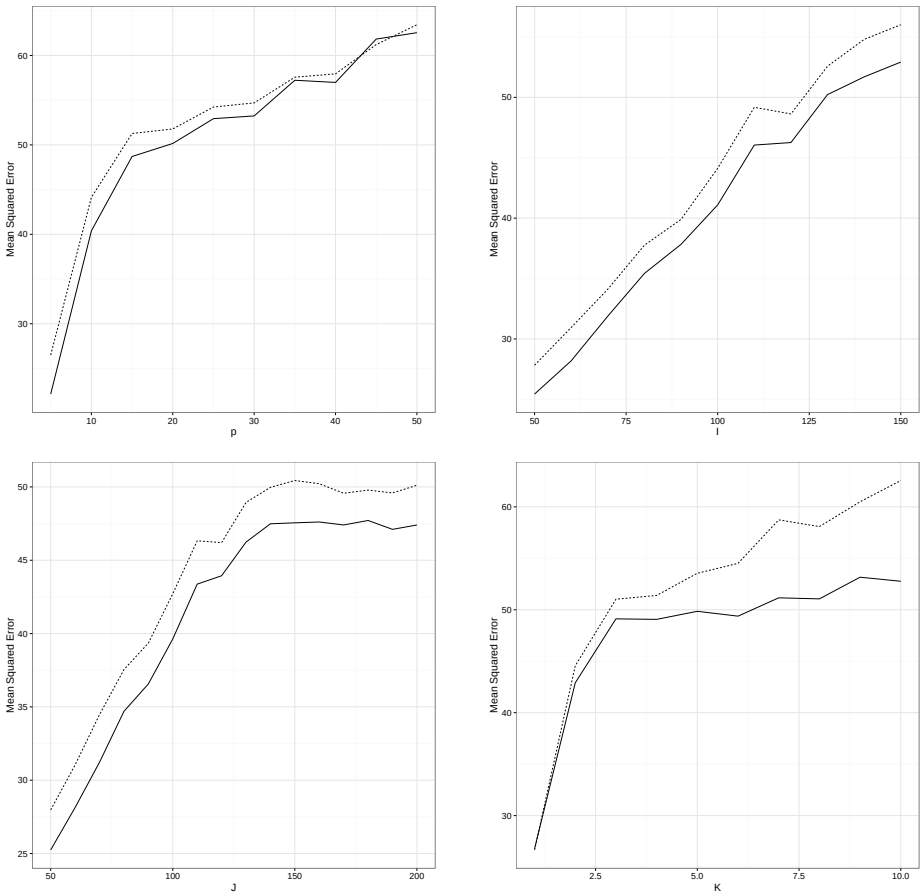


Figure 2.7: Comparison of the estimate of posterior mean among repeated observations from functional data generated from Fourier basis family with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

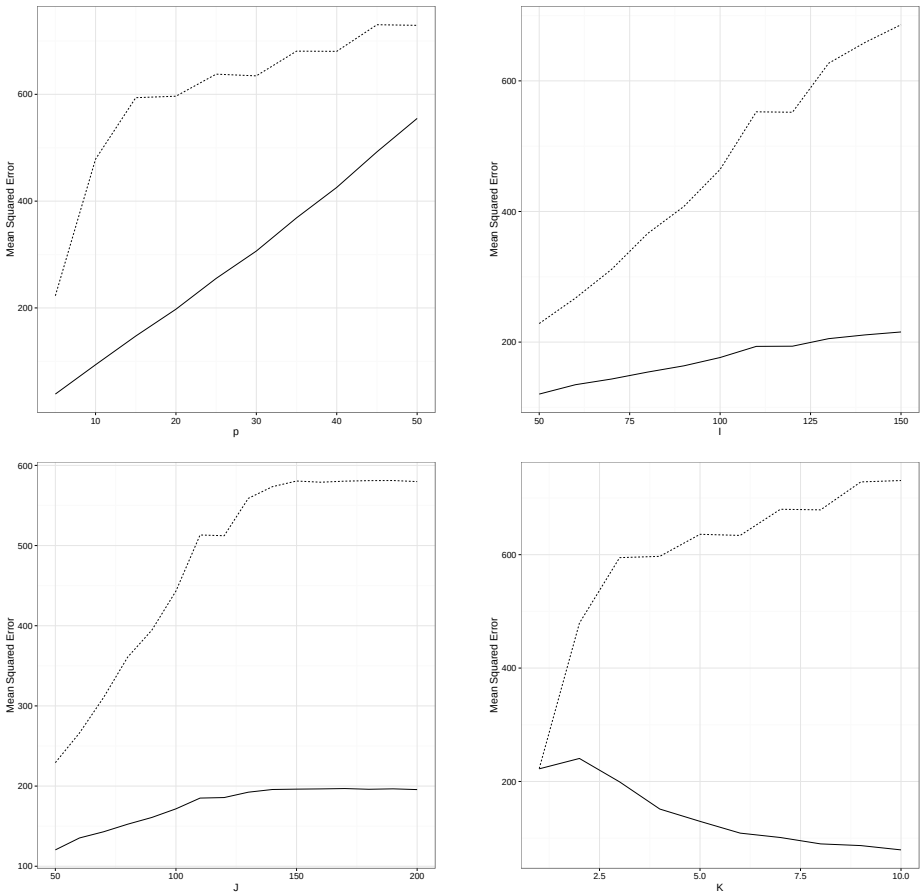


Figure 2.8: Comparison of the estimate of posterior global variance among repeated observations from functional data generated from Fourier basis family with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

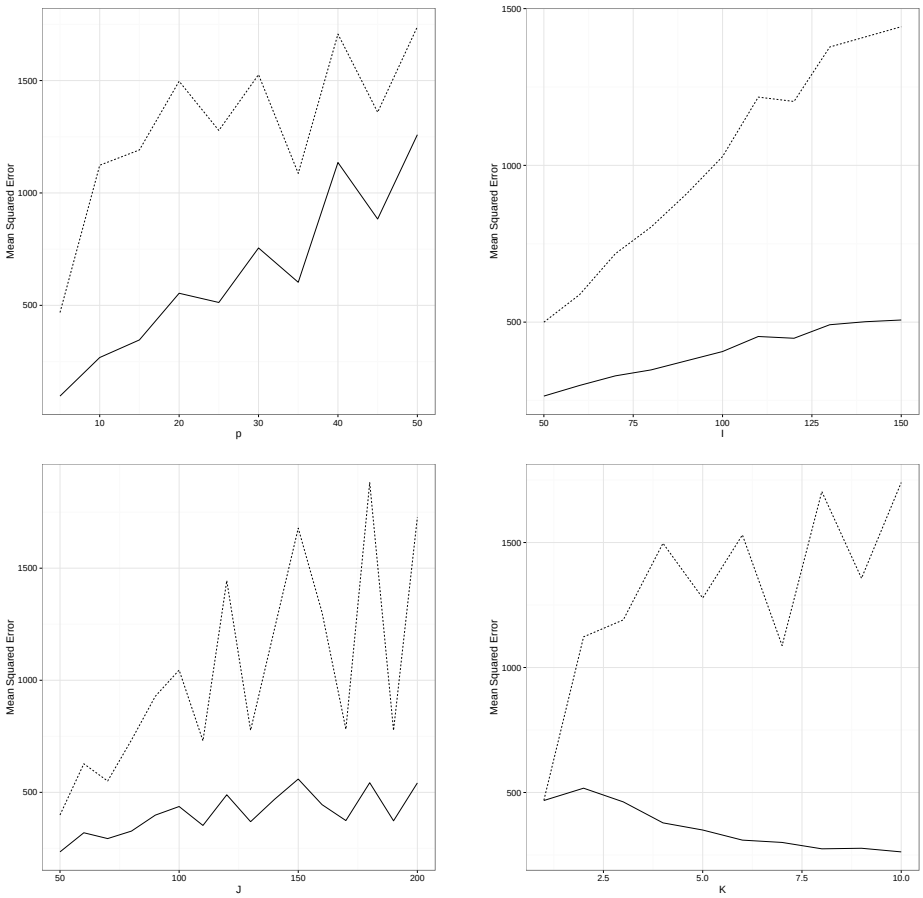


Figure 2.9: Comparison of the estimate of posterior variance among repeated observations from functional data generated from Fourier basis family with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

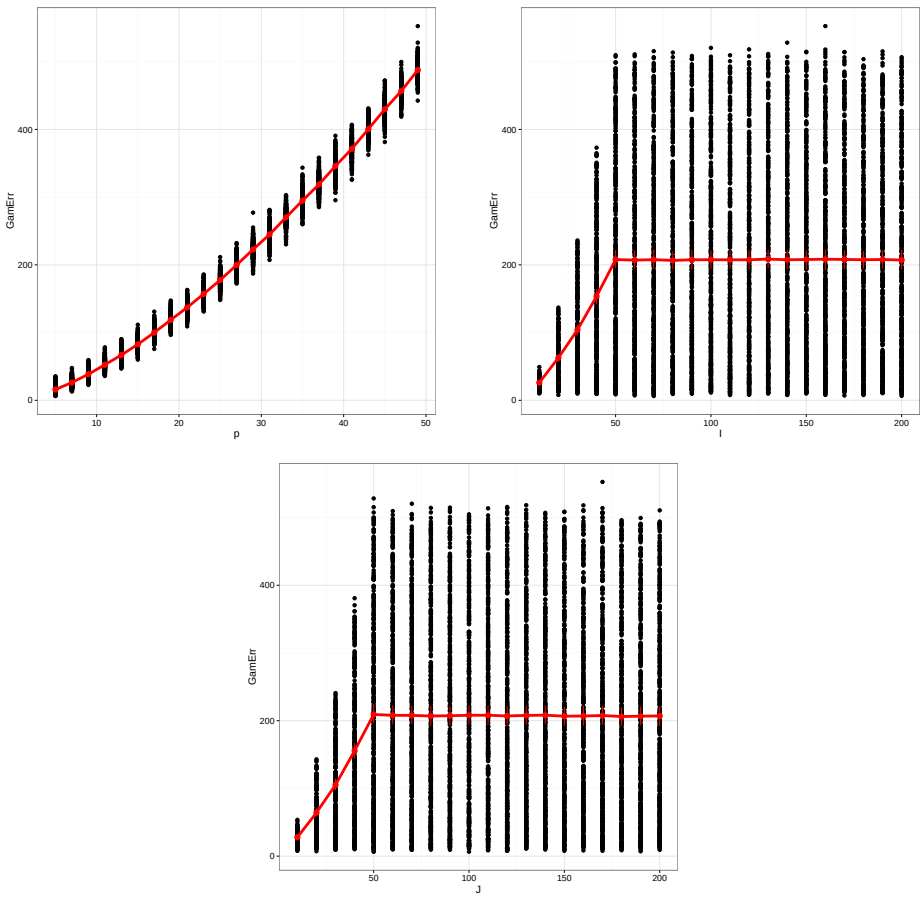


Figure 2.10: Comparison of the estimate of posterior variance among individuals from functional data generated from Fourier basis family with (a) changing dimension and (b) changing number of individuals and (c) changing number of observations per individual.

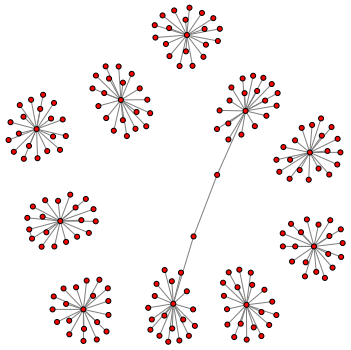


Figure 2.11: Visualization of graphical structure of the errors

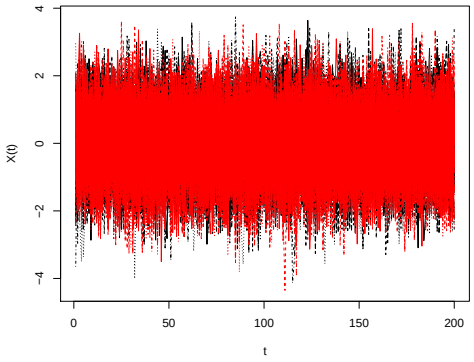


Figure 2.12: Two classes generated from different covariance structure but similar means

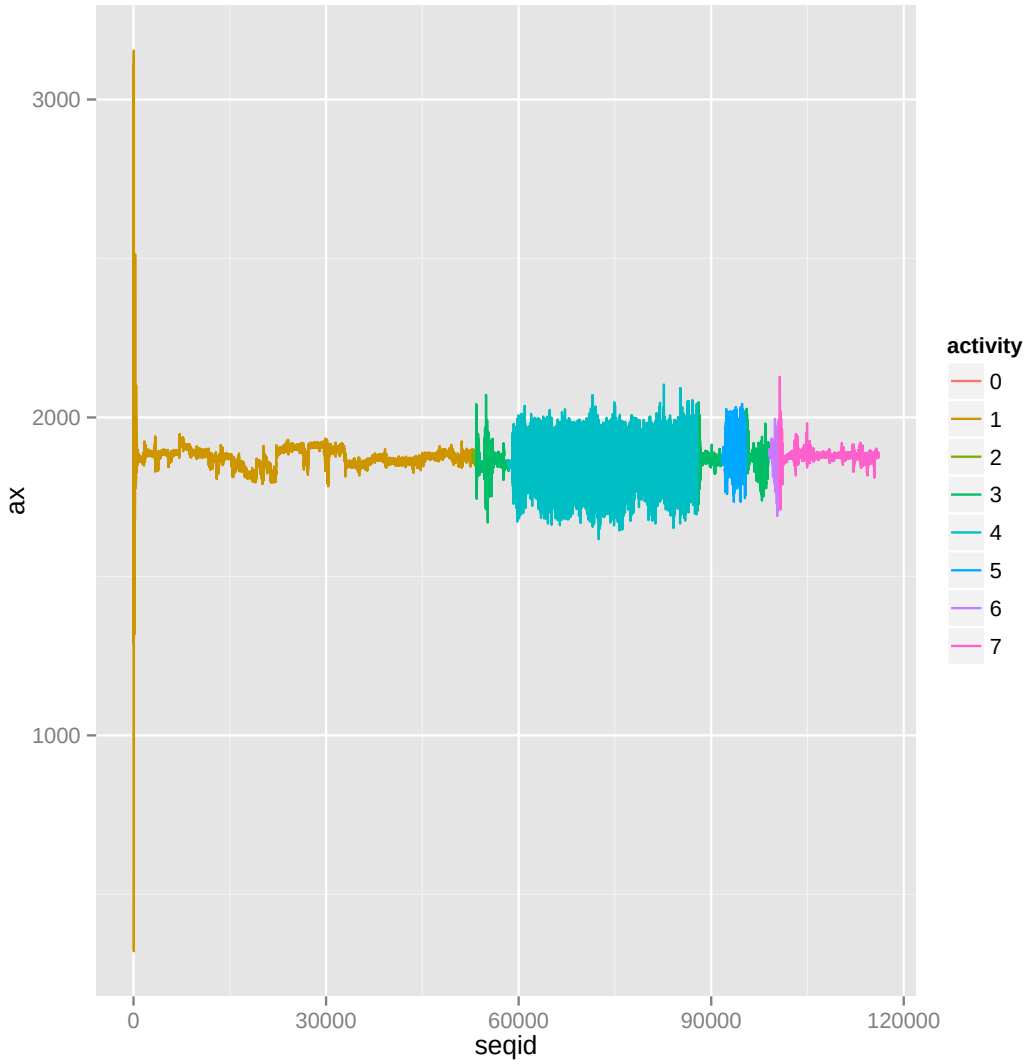


Figure 2.13: The X acceleration of the subject 14

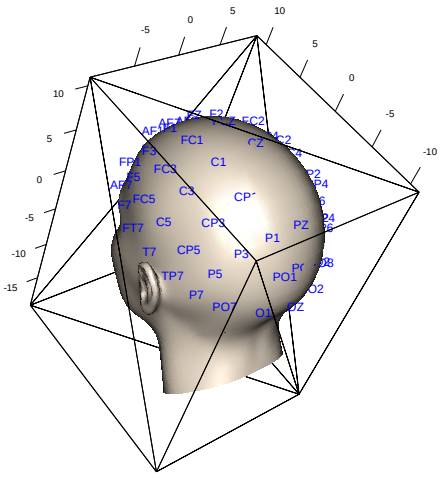


Figure 2.14: Position of the electrodes used in the dataset

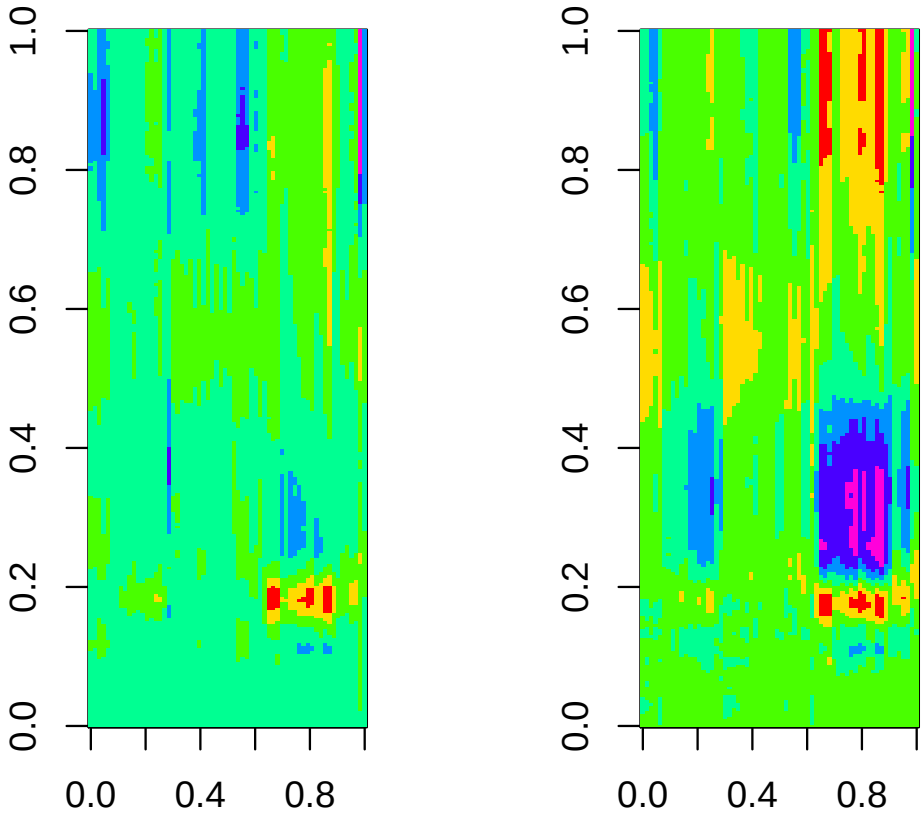


Figure 2.15: Mean response for Stimuli 1 (Single picture shown). Left panel shows alcoholics and right panel shows control. Time is at the horizontal axis and channels are at vertical axis.

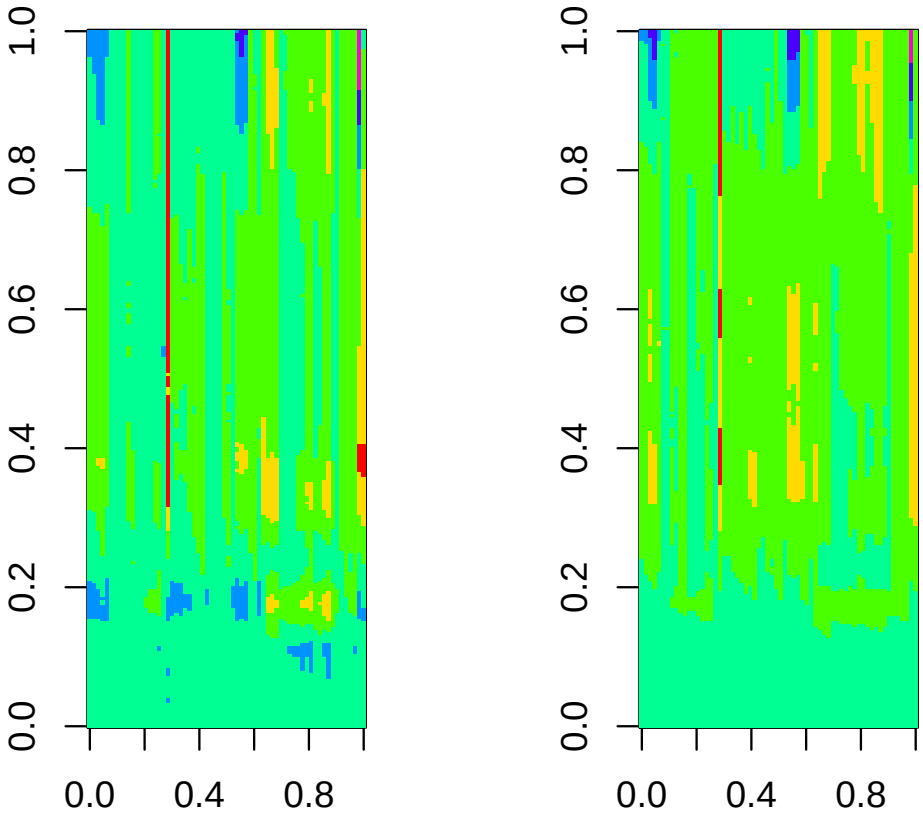


Figure 2.16: Mean response for Stimuli 2 Matching (Two matching pictures shown).

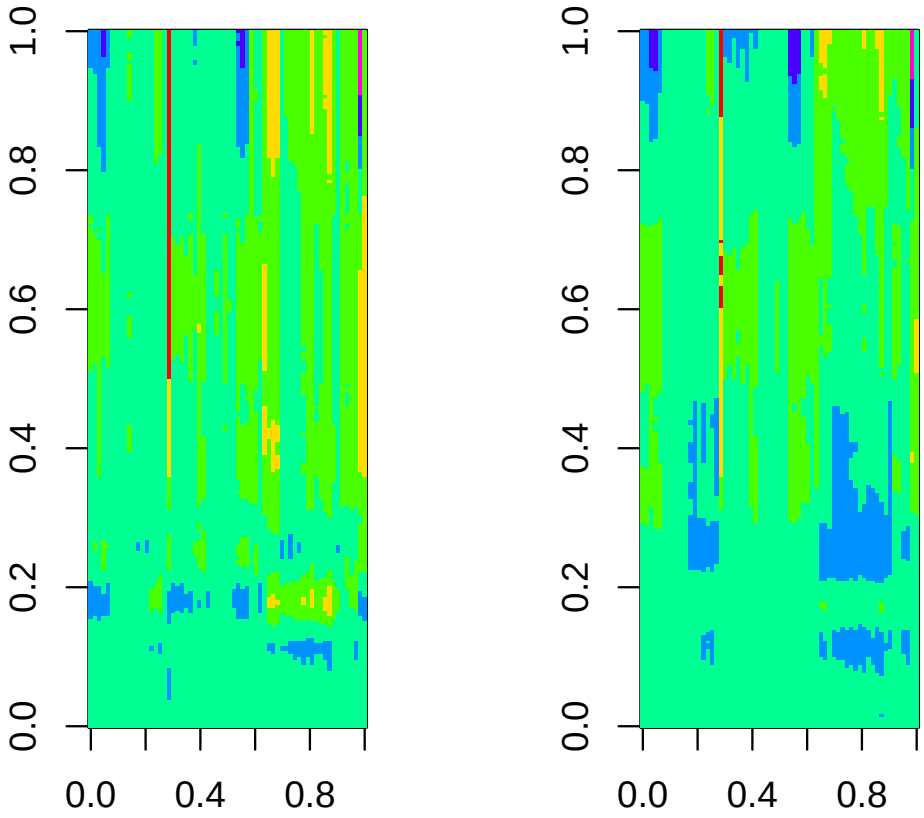


Figure 2.17: Mean response for Stimuli 2 Non-matching (Two non matching pictures shown.)

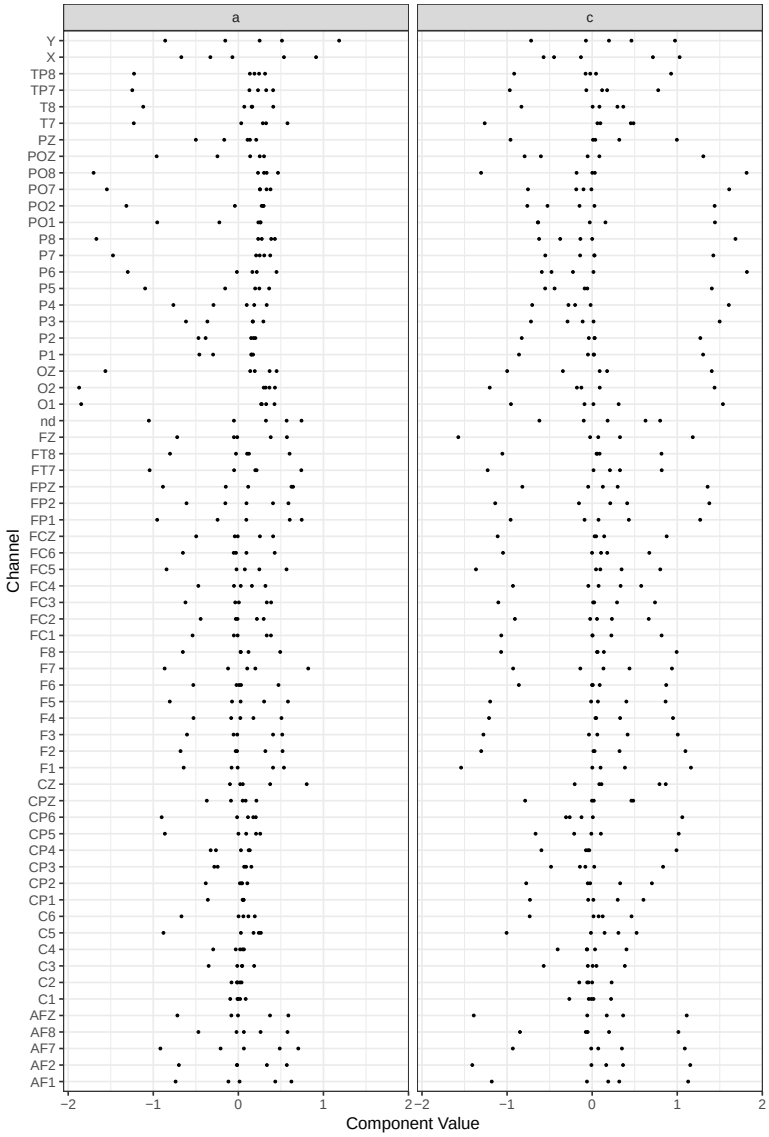


Figure 2.18: Channel wise components of EEG data for two groups.

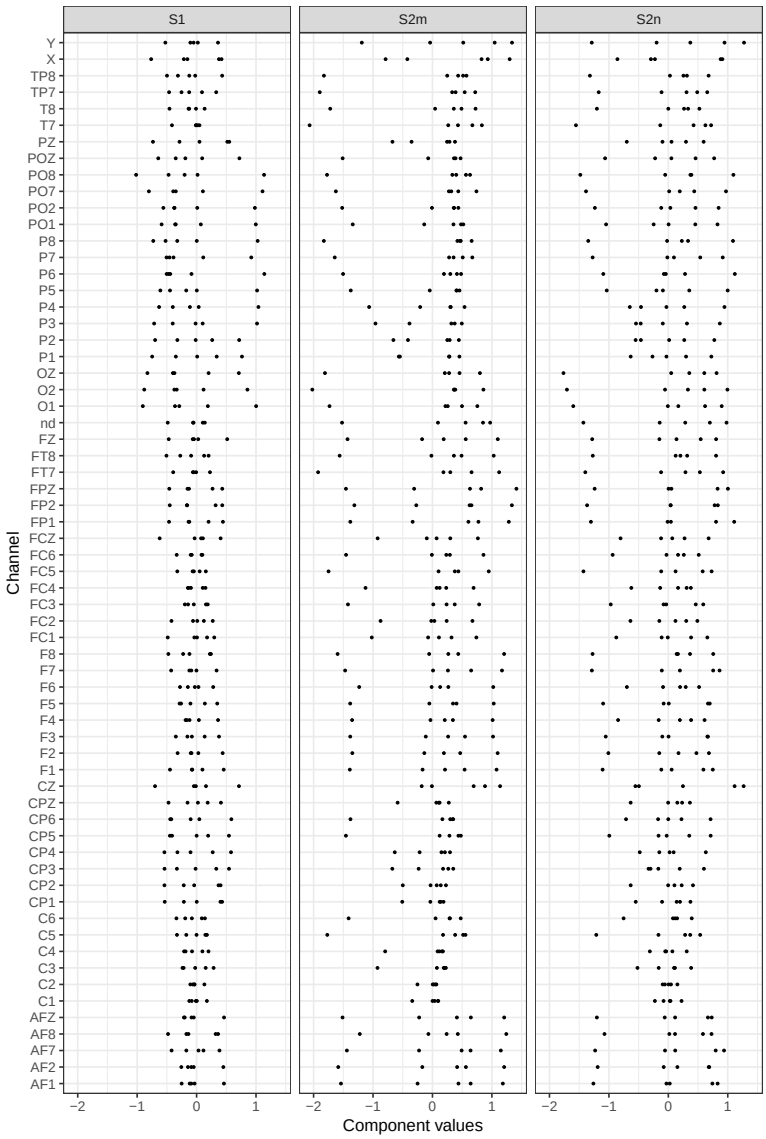


Figure 2.19: Channel wise components of EEG data for three kinds of stimulus.

Chapter 3

A Bootstrap Based Multiple Hypothesis Testing Procedure

3.1 Introduction

Multiple hypothesis testing is a classical problem in statistics. In recent years, this problem has gained relevance due to a variety of applications in research in medical fields. In these areas of research, sometimes conclusions are drawn by simultaneous testing of a large number of hypothesis, which is the problem of consideration in multiple hypothesis testing. In these situations, a single testing approach is known to fail. Hence, there is a need for controlling the false discoveries while maximizing power of each individual tests. There has been various criteria of rate control been proposed, and there are multiple approaches available. Use of multiple hypothesis testing is very common in Brain imaging for identifying activity of neurons in the brain Worsley et al. [40], Ellis et al. [9], Merriam and Genovese [24]. Imaging techniques like functional FMRI use it too Logan [23]. It is also used for identifying difference in gene expression in micro-array experiments Drigalenko [6], Weller and Song [37, 37], Heyen et al. [15], Bovenhuis [3], Mosig et al. [25], Reiner et al. [29]. Dudoit and Shaffer [8], Sebastiani et al. [32] provides a comprehensive review of multiple hypothesis testing in gene expression applications. There are examples of application of multiple hypothesis testing in medicine Khatri and Babyyak [21], Schlaeppi and Edwards [31], public health studies Ottenbacher [26], Vedantham et al. [36], marketing Schaffer [30].

As a problem, multiple hypothesis testing is essential in several areas of statistics, and many other problems can be solved using it. Variable selection George [13, 13], item response theory Ip [18], structural equation modeling George [13],

decision trees Yekutieli et al. [41], Abramovich and Benjamini [1] etc. has applications of multiple hypothesis testing. We will talk about the statistical applications of multiple hypothesis testing in more details.

Many of the applications mentioned above uses functional data or has their equivalent in the functional data world. For example, brain imaging data can be considered functional. In fact for any multi-channel functional data, the problem of multiple hypothesis testing could arise. We first describe the problem formally. We consider m tests are being performed. The outcome of a single test can be one of the following: the number of rejection $R = N_{1|0} + N_{1|1}$ the sum of true rejections and false rejections. Usually, for a single test, the test statistics cut off is chosen when probability of false rejection or type I error is controlled, while maximizing the power of the test. But for multiple tests, having the same level of type I error is not possible as the error accumulated from each test could get large. In terms of p values, it becomes more likely for the p values to be smaller when large number of them calculated. In these situations, need for an appropriate correction becomes important.

The control of type I error could be implemented in a variety of measures, as there are multiple definitions of measures with slightly different ideas. The control on type I error means the error rate is kept bound within a $\alpha\ in(0, 1)$. The most common measure used here is Family-wise error rate(FWER). It is conditional on only Null hypothesis.

$$FWER = P(N_{1|0} \geq 1)$$

There are several methods for controlling FWER Holm [16], discusses the situation when the relative importance of the single hypothesis are known. Westfall et al. [39] provides a way for generalizing in high dimensional situations.

Several other measures of type I error are based on False discovery proportion (FDP).

$$FDP = \begin{cases} \frac{N_{1|0}}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

In the approach of Benjamini and Hochberg Yoav Benjamini [42] referred to expectation of FDP as False Discovery Rate. Dudoit Dudoit [7] and Genovese and Wasserman [11] propose to control the tail probability of FDP, which is sometimes referred to as False discovery exceedance.

$$FDR = E[FDP]$$

$$tFDP(c) = P[FDP > c]$$

These definitions has been generalized for various situations Schaffer [30], Westfall [38], Dudoit [7]. Storey Schaffer [30] introduced "positive FDR" defined as pFDR = $E[FDP|R > 0]$. Storey suggested the estimation and control of pFDR fixed rejection region, and introduced q-value, the p-FDR analogue of the p-value. Later he also provided a Bayesian interpretation of q-values JD [20], which also points to a connection with classification theory. Weighted control of FDR was proposed in Benjamini and Hochberg [2] and Genovese et al. [12].

There are several different approaches available for implementing the ideas mentioned above. Augmentation van der Laan MJ [35], inversion Genovese and

Wasserman [11] or bootstrap based Pesarin [27] are some of the most commonly used approaches.

There are instances of use of multiple hypothesis testing in medical imaging. We wish to use it in the context of classification. The problem of classification from multi-channel data could be transformed into a series of multiple hypothesis testing problem. The relative importance of the channel in discriminating between classes is of importance to us, hence we propose use a hypothesis testing on the channels to eliminate the ones which are less important. Next, on the smaller number of channels, our proposed method of hierarchical classification could be used.

3.2 Experiment with different dependence structure

In this experiment, we generated data from multivariate normal $X \sim N_m(\mu, \Sigma)$. We used μ to be a vector of all zero except for m_0 number of co-ordinates. Those m_0 co-ordinates were chosen randomly and set to 1. We then performed co-ordinate wise t-tests and collected p-values and used our method to get the significant cut-off point. For this experiment, we set $m = 1000$, $m_0 = 10$. We observe $n = 20$ observations of X. We use different Σ to generate the data. The forms of $Sigma$ were: $\sigma^2 * I_m$, Autocorrelated decay, and Compound symmetry. With each of the setup, the marginal variances were $\{0.5, 0.75, 1, 1.25\}$. The ρ parameter for auto correlated decay was chosen to be $\rho = 0.25$. The compound symmetry

additive parameter was chosen to be $\sigma_1^2 = 0.25$.

3.3 Simulation Study on Various Parameters

To build intuition on our method, we set up the following experiments:

3.3.1 Effect of Sample Size

We used sample size = $\{20, 50, 100, 200\}$ and the Σ was chosen as the identity matrix. Dimension was $m = 1000$, with $m_0 = 10$, signal size = 1, $\sigma^2 = 0.5$.

In figure 3.3.1, we show the effect of sample size increases for bootstrap and in figure 3.3.1, we show the effect of independent sample on the distribution of the order statistics of the p values.

Looking at the pictures, it seems the bootstrap elbow point does not get affected by the sample size as much. The quantile curves gets closer to each other. However for the independent case, the elbow point becomes sharper from 20 to 50. But after that the improvement is marginal. The smoothness of the curves do not get affected by the sample size, where, in bootstrap case, that is visible.

3.4 Methodology

We propose a new methodology for multiple hypothesis testing based on bootstrap distribution of p values. We know that under null hypothesis p value is an uniform$(0, 1)$ random variable. If the hypothesis are independent, then so are the

p values. We use this intuition to construct a statistic that can be used to test multiple hypothesis.

We consider the case of coordinate wise testing of a data vector. If the data matrix X is of dimension $I \times J$ then without loss of generality, we consider the following set of hypothesis for $j = 1, 2, \cdots, J$.

$$H_0^{(j)} : x_j = 0$$
$$H_1^{(j)} : x_j \neq 0$$

Where x_j denotes the jth coordinate of X. This set of tests can be performed in a variety of ways, for the purpose of illustration, we choose to use one sample t-tests.

The general procedure for testing multiple hypothesis is testing independently then use a correction method to correct for family wise error rate. Here we are not using any such correction, as we are using a bootstrap based approach.

As we are dealing with distribution of p values, we can show some elementary results about the distributions of quantities related to p values when the null hypothesis is true. The following two results follow from distribution theory and distribution of order statistics.

Lemma 6. *If $p_{(j)}$ denotes the jth order statistics of p values from a set of hypothesis $\{H_1, \cdots, H_n\}$, which are independent from each other. We also assume for $i = 1, \cdots, n$, the null hypothesis are true. Then the order statics of the p values follow the distribution*

$$p_{(j)} \sim Beta(j, (n + 1 - j)) \tag{3.1}$$

Proof. We first look at the distribution of the p values when null is true and the tests are independent. We then get the order statistics of them.

Let T_j be the test statistic for hypothesis H_j, with rejection region R_j. The following is how p values are generated.

$$P(P_j < p_j) = P(T_j \in R_j) = p_j \tag{3.2}$$

Hence, the p values follow a uniform distribution, when null hypothesis is true.

$$p_j \sim_{iid} U(0,1) \tag{3.3}$$

So, order statistics of the p values can be written using the following equation

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^(j-1)[1-F_X(x)](n-j) \tag{3.4}$$

As $f_{P_j}(p_j) = p_j$ for all $j = \{1, \cdots, n\}$, we can write the distribution of $p_{(j)}$ as

$$f_{P_{(j)}}(p) = \frac{n!}{(j-1)!(n-j)!} p^{(j-1)}(1-p)^{(n-j)} \tag{3.5}$$

Hence, the jth order statistic of the p values follow a beta distribution with shape j and scale parameter $(n+1-j)$.

□

Next, we deal with the difference of the consecutive order statistics of the p values. This will be used later in our methodology as difference is a very common operator for change detection in time series. We consider a monotone transformation of the order statistics of the p values for convenience.

Lemma 7. *We define the following quantities.*

$$\begin{aligned} Z_0 &= 0 && (3.6)\\ Z_j &= -\log(1-p_j) && (3.7)\\ W_1 &= Z_{(1)} && (3.8)\\ W_j &= (Z_{(j)} - Z_{(j-1)}) && (3.9)\\ & && (3.10) \end{aligned}$$

Where $Z_{(j)}$ denotes the jth order statistic of Z. When the null hypothesis $\{H_1, \cdots, H_n\}$ are all true and independent, then we can say the following about the distributions of $Z_{(j)}$ and W_j

$$\begin{aligned} Z_{(j)} &\sim_{ind} Exp(1/n) && (3.11)\\ W_j &\sim_{ind} Exp((n+1-j)^{-1}) && (3.12) \end{aligned}$$

Proof. Let us define the vectors Z and W as follows

$$\begin{aligned} Z &= (Z_{(1)}, \cdots, Z_{(n)}) && (3.13)\\ W &= (W_1, \cdots, W_n) && (3.14) \end{aligned}$$

We first write the relationship between W and Z variables as a linear transformation. For a matrix M, we can write

$$W = MZ \tag{3.15}$$

where, the matrix of transformation is given by

$$M = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{3.16}$$

So, we can write Z in terms of W as

$$Z = M^{-1}W \tag{3.17}$$

We know that M^{-1} exists as, M is a lower triangular matrix with non zero diagonals. Hence,

$$\begin{aligned} det(M) &= \prod_{i=1}^{n}[m_{ii}] && (3.18) \\ &= 1 && (3.19) \end{aligned}$$

So, the distribution of W can be written in terms of Z as follows, using the value of the determinant from above and using the variable transformation method.

$$\begin{aligned} f_W(w) &= f_Z(z)(M^{-1}w)\frac{1}{|det(M)|} && (3.20) \\ &= f_Z(M^{-1}w) && (3.21) \\ & && (3.22) \end{aligned}$$

The joint distribution $\{Z_{(1)}, \cdots, Z_{(n)}\}$ is given by the following result about all order statistic from a sample

$$f_{Z_{(1)},\cdots,Z_{(n)}} = n! \prod_{j=1}^{n} f(Z_j) \text{ if } Z_1 < \cdots < Z_n \tag{3.23}$$

As, $(w_1 < w_1 + w_2 < \cdots < w_1 + \cdots + w_n)$ we write from equation (3.20) as follows

$$\begin{aligned}
f_W(w) &= f_Z(M^{-1}w) && (3.24)\\
&= n! f_W(w_1) f(w_1 + w_2) \cdots f(w_1 + \cdots + w_n) && (3.25)\\
&= n! e^{-y_{(}1)} e^{-(y_{(}1)+y_{(}2))} \cdots e^{-(y_{(}1)+\cdots+y_{(}n))} && (3.26)\\
&= n! \exp(-n(y_{(1)}) - (n-1)y_{(2)} \cdots y_{(n)}) && (3.27)\\
&= n! \exp(-ny_{(1)}) \exp(-(n-1)y_{(2)}) \cdots \exp(y_{(n)}) && (3.28)\\
&= (n \exp(-ny_{(1)}))(n-1) \exp(-(n-1)(y_{(2)} - y_{(1)})) \cdots \exp(y_{(n)} - y_{(n-1)}) && (3.29)\\
&= \prod_{j=1}^{n} (n+1-j) \exp(-(n+1-j)(y_{(j)} - y_{(j-1)}) && (3.30)
\end{aligned}$$

As we know that the the hypothesis are independent hence the p values and other related statistics are all independent. Hence, each coordinate of the vector W has the exponential distribution given in equation 3.11.

The same result can also be argued using the memory less property of exponential distribution.

□

We first generate bootstrap samples of our data, let us denote the bootstrap

samples by $X^{\{b\}}$ for $b \in \{1, 2, \cdots, B\}$, where B is the bootstrap sample size. For each of these samples we can perform J tests in parallel and collect the p values. We denote the p values from jth coordinate of bth bootstrap sample by $p_j^{\{b\}}$. We then use a monotone transformation of the p values for better visualizing. The transformed order statistics are collected as shown.

$$Z_{(0)}^{\{b\}} \stackrel{\text{def}}{=} 0 \tag{3.31}$$

$$Z_{(j)}^{\{b\}} = -\log(1 - p_{(j)}^{\{b\}}) \tag{3.32}$$

$$W_j^{\{b\}} \stackrel{\text{def}}{=} Z_{(j)}^{\{b\}} - Z_{(j-1)}^{\{b\}} - (n + 1 - j)^{-1} \tag{3.33}$$

With this simple formulation, we can easily check that a linear transformation of W_j is exponentially distributed.

However, the distribution of W_j under alternative is not so simple and would depend on the specification of the alternative. But, if we are interested in detecting the split between the tests from where alternative becomes true, we have to detect a change in the distribution of W. We will try to explain this idea and this will be implemented in a few different ways.

Firstly, we consider the order statistics $p_{(j)}$ themselves. Let us define the quantile function $Q(X, q) = \min_x\{P(X \leq x) \geq q\}$ for $q \in (0, 1)$.

3.4.1 Change in Cumulative Sum

Let J be the number of hypothesis, indexed by $\{H_1, \cdots, H_J\}$. Without loss of generality, we are assuming that for the first k hypothesis, the null hypothesis is

true. As we are considering the order statistic of the empirical distribution of the p values, we can expect then the quantiles of the $\{p_{(1)}, \cdots, p_{(k)}\}$, should be very close to zero. Also, $Q(p_{(j)}, q) \geq Q(p_{(j-1)}, q)$ for all q. So, if we take the cumulative sum for the left part of a point, it must be lower than the cumulative sum from the right part of the point. We divide by the number of points to keep the scale fixed. This gives rise to the following two statistics.

$$CML(j) = \frac{1}{j}\sum_{i=1}^{j} Q(p_{(i)}, q) \tag{3.34}$$

$$CMR(j) = \frac{1}{j}\sum_{i=j}^{n} Q(p_{(i)}, q) \tag{3.35}$$

Based on the two cumulative sum statistics, we estimate the value of k by the point with maximum difference.

$$\hat{K}_{CM} = arg\max_{j}(CMR(j) - CML(j)) \tag{3.36}$$

3.4.2 Change in Linear Indexing

The shape of the curve of $p_{(j)}$ as a function of j should be monotonically increasing. However, as we have assumed that the first k hypothesis, the null is true, so, we hope to see a sharp change in the relationship of $p_{(j)}$ with the index after k. So, if we fit a linear model on the left part of the curve of k and the right part of k, we should see a large difference in the slope parameter. Following two statistics try to capture these two slopes.

$$LM(j) = arg\min_{\beta} \sum_{i=1}^{j} ||Q(p_{(i)}, q) - i\beta||^2 \tag{3.37}$$

$$RM(j) = arg\min_{\beta} \sum_{i=j}^{n} ||Q(p_{(i)}, q) - i\beta||^2 \tag{3.38}$$

Similarly, we can estimate the change point in the curve of $Q(., q)$ with the change statistic constructed using the above definitions.

$$\hat{K}_{LM} = arg\max_{j} |RM(j) - LM(j)| \tag{3.39}$$

3.4.3 Local Linear Regression

Here we use similar idea as of linear indexing, however, we try to model the shape of the curve using a local linear regression method. For a detailed study of local linear regression method we refer to Loader [22]. Here we present a brief review of the method for our purpose.

With a data set $(x_1, Y_1), \cdots, (x_n, Y_n)$, where the predictor is $(x_1, \cdots, x_n)$ and response Y, the relationship is assumed to be a non linear model given in equation (3.40).

$$Y_i = \mu(x_i) + \epsilon_i \tag{3.40}$$

The errors ϵ_i are assumed to be iid with mean $E(\epsilon_i) = 0$ and finite variance $E(\epsilon_i^2) = \sigma^2 < \infty$.

The difference between this method of regression with a model based method

is that the link function $\mu(.)$ is not assumed to belong to any parametric family. The fitting procedure uses a smoothing window based method.

For a fixed point x, the bandwidth $h(x)$ and a smoothing window $(x-h(x), x+h(x))$ is defined. The predicted value of the response is a weighted average of the response in this window. The weights are given by a weighting function in equation (3.41).

$$w_i(x) = W\left(\frac{x_i - x}{h(x)}\right) \tag{3.41}$$

The weighting function $W(.)$ is any function that typically assigns less weight to far away points from x. However the weighting scheme can be different between different weighting functions. This is similar to the concept of kernels prevalent in non parametric statistics as any Kernel function (for example NW kernel) would work here. We are using the R package *locfit* for local regression, which by default chooses a tri cube kernel.

Finally, when we use the same approach of fitting model using the left and right part of the data. However, here we do not have a parametric model of the curve, so we decide to use the difference of the predicted values to create the metric for change detection. We describe this statistic in equation (3.42), where the $\hat{Q}$ denotes the predicted value of the quantile function. The difference statistic is given in equation (3.44).

$$LLL(j) = \sum_{i=1}^{j} \hat{Q}(p_{(i)}, q) \tag{3.42}$$

$$RLL(j) = \sum_{i=j}^{n} \hat{Q}(p_{(i)}, q) \tag{3.43}$$

$$\hat{K}_{LR} = arg\max_{j}(RLL(j) - LLL(j)) \tag{3.44}$$

3.4.4 Comparison of Change Detection Methods

We perform a simple simulation study to compare the performance of the three change detection methods described above. We generate data from three scenarios, that resemble the shape of the quantile plot of the order statistics of the p values from real hypothesis tests. Let the set of the test statistics be $\{y_1, \cdots, y_n\}$, with a true change point at y_k. In connection with our terminology, these will be $Q(p_i, q)$, however for simplicity of notation, we use y.

1. Constant to linear change: $\{y_1, \cdots, y_k = 0\}$ and $y_{k+1}, \cdots, y_n$ are linearly increasing.

2. Linear to linear change: $\{y_1, \cdots, y_k\}$ are increasing with a slope β_1 and $y_{k+1}, \cdots, y_n$ are increasing with a slope β_2, with $\beta_2 > \beta_1$.

3. Linear to quadratic change: $\{y_1, \cdots, y_k\}$ are increasing and $y_{k+1}, \cdots, y_n$ are increasing quadratically.

These scenarios are designed based on real hypothesis test cases, however, the nature of the curve after the break point is still subject to assumption, as it depends on the specification of the alternative hypothesis. In figure 3.3, the performances of the three change detection methods are shown for each data generation scenarios. It is clear that for constant linear and linear linear change, all three methods perform almost equally well. However, for the linear quadratic change, the method of local linear regression performed well. This is expected as other two methods depend on the linearity of the statistics. Even though, for the non linear case, the method of local regression performed well, in real experiments, we have found that it is possible to outperform it using appropriate truncation of the data with other two methods.

3.5 Theoretical Properties

We assume that under the j^{th} null hypothesis, the test statistic T_j has a distribution that converges to a standard normal distribution $N(0,1)$ as the sample size n tends to ∞. We also assume that under the alternative H_{1j}, for some non-zero constant δ_j, the distribution of $T_j - n^{1/2}\delta_j$ has the weak limit of the standard normal distribution. These assumptions rarely fail to hold, and can be easily shown to be valid for very standard hypothesis tests like the one, two or paired sample t-tests, distribution free tests, and more complex cases. Also, this assumption can be relaxed considerably for the results presented here: we need to assume only that under H_{0j} and H_{1j} respectively the random variates T_j and $T_j - a_n\delta_j$ have

some limiting distribution, where a_n is some sequence tending to infinity with the sample size. We do not present the results under this more general case to avoid lengthy mathematical details. We adopt the notation that T_{0j} is a random variable that is independent of all data and consequently independent of all the test statistics, and has the same distribution as that of T_j under H_{0j}, thus its distribution limits to the standard normal with increasing sample size.

For the moment, we make the major assumption that the p-values P_j are independent random variables. This assumption is not required for the algorithm and methodology presented in this paper. However, the analysis under the assumption of independence brings clarity in the theoretical developments that underlie the methodology we propose in this paper. Without loss of generality, $0 < |\delta_1| \leq \ldots \leq |\delta_{p_1}|$. Note that we allow for one or more of the δ values to be equal. For the sake of notational completeness, we define $\delta_j = 0$ when H_{0j} is true, and define the indicator function $\mathcal{I}_{H_{1j}}$ to be 0 if H_{0j} is true and to be one otherwise. We consider a scenario where a high absolute value of the test statistic T_j demonstrates lack of compatibility of the data with the null hypothesis H_{0j}. This implies that the j^{th} p-value P_j is $P_j = \mathbb{P}\big[|T_{0j}| \geq |t_j| \big| T_j = t_j\big]$, which under our assumptions asymptote as $2\overline{\Phi}(|Z_j + n^{1/2}\delta_j \mathcal{I}_{H_{1j}}|)$, where Z_j is a standard normal random variable.

Thus, the p-values that correspond to the coordinates where the null hypothesis is false are $\tilde{U}_j = 2\overline{\Phi}(|Z_j + n^{1/2}\delta_j|)$, $j = 1, \ldots, p_1$ while those that correspond to coordinates where the null hypothesis is true are $U_j \sim Uniform(0, 1)$, $j = 1, \ldots, p - p_1$. Suppose we arrange the $\tilde{U}_j$'s in an ascending order, thus $\tilde{U}_{(1)} <$

$\ldots < \tilde{U}_{(p_1-1)} < \tilde{U}_{(p_1)}$. Similarly, $U_{(1)} < \ldots < U_{(p-p_1)}$.

Our proposal effectively utilizes the following fact:

Theorem 3.5.1. *Suppose p and p_1 are fixed integers, and $\delta_1, \ldots, \delta_{p_1}$ fixed non-zero real numbers. Then as sample size n tends to infinity, with a probability $1 - O(\exp\{-Cn\})$ for some constant $C > 0$, we have the following relationship:*

$$\tilde{U}_{(1)} < \ldots < \tilde{U}_{(p_1-1)} < \tilde{U}_{(p_1)} \ < \ U_{(1)} < \ldots < U_{(p-p_1)},$$

and the maximum expected second difference between successive terms in this sequence occurs at position p_1.

Proof of Theorem 3.5.1. First, notice that the only result to establish for the first part is that $U_{(1)} - \tilde{U}_{(p_1)} > 0$ with probability $1 - O(\exp\{-Cn\})$. When we prove the second part that this gap is the highest among all differences between succesive terms, this result follows. Note that $\mathbb{E}U_{(j)} = j/(p - p_1 + 1)$. This implies that the expected gap between $U_{(j+1)}$ and U_j is $1/(p - p_1 + 1)$.

We will now show that with probability $1 - O(\exp\{-Cn\})$,

$$\tilde{U}_j < 4n^{-1/2}|\delta_j^{-1}|\phi(Z_j + n^{1/2}\delta_j).$$

Consider the sets

$$\begin{aligned}
\mathcal{A}_0 &= \Big\{U_{(1)} > 4n^{-1/2}|\delta_1^{-1}|\phi(n^{1/2}\delta_1)\Big\}, \\
\mathcal{A}_j &= \Big\{|Z_j| < n^{1/2}|\delta_j|/2\Big\}, \quad j = 1, \ldots, p_1, \\
\mathcal{B} &= \mathcal{A}_0 \cap \mathcal{A}_1 \ldots \mathcal{A}_{p_1}.
\end{aligned}$$

We first establish that in the set $\mathcal{A}_j$, we have

$$|Z_j + n^{1/2}\delta_j| > n^{1/2}|\delta_j|/2.$$

Suppose that $\delta_j > 0$. Then on the set $\mathcal{A}_j$, we have $Z_j > -n^{1/2}\delta_j/2$. Then

$$|Z_j + n^{1/2}\delta_j| \geq Z_j + n^{1/2}\delta_j > n^{1/2}\delta_j/2 = n^{1/2}|\delta_j|/2.$$

Suppose that $\delta_j < 0$. Then on the set $\mathcal{A}_j$, we have $Z_j < -n^{1/2}\delta_j/2$. Then

$$Z_j + n^{1/2}\delta_j < n^{1/2}\delta_j/2 < 0.$$

Consequently, $|Z_j + n^{1/2}\delta_j| > n^{1/2}|\delta_j|/2$.

Now notice that on the set $\mathcal{A}_j$, we have

$$\begin{aligned} 2\overline{\Phi}(|Z_j + n^{1/2}\delta_j|) &\leq 2|Z_j + n^{1/2}\delta_j|^{-1}\phi(Z_j + n^{1/2}\delta_j) \\ &\leq 4n^{-1/2}|\delta_j^{-1}|\phi(Z_j + n^{1/2}\delta_j). \end{aligned}$$

Also on this set, we have

$$\phi(Z_j + n^{1/2}\delta_j) = \phi(|Z_j + n^{1/2}\delta_j|) < \phi(|n^{1/2}\delta_j|)$$

since for $x > 0$, $\phi(x)$ is a decreasing function.

Consequently, on the set $\mathcal{A}_j$, we have

$$2\overline{\Phi}(|Z_j + n^{1/2}\delta_j|) \leq 4n^{-1/2}|\delta_j^{-1}|\phi(|n^{1/2}\delta_j|) \leq 4n^{-1/2}|\delta_1^{-1}|\phi(|n^{1/2}\delta_1|).$$

Also notice that by definition, on the set $\mathcal{A}_0$, $U_{(1)} > 4n^{-1/2}|\delta_1^{-1}|\phi(|n^{1/2}\delta_1|)$.

Consequently on the set $\mathcal{B}$, our first result is established. We now need to show that the probability of this set is $1 - O(\exp\{-Cn\})$.

First, note that

$$\mathbb{P}\Big[U_{(1)} > u\Big] = \Big(1-u\Big)^{p-p_1},$$

hence

$$\begin{aligned}\mathbb{P}\Big[U_{(1)} > 4n^{-1/2}|\delta_1^{-1}|\phi(|n^{1/2}\delta_1|)\Big] &= \Big(1-4n^{-1/2}|\delta_1^{-1}|\phi(|n^{1/2}\delta_1|)\Big)^{p-p_1} \\ &> 1-4(p-p_1)n^{-1/2}|\delta_1^{-1}|\phi(|n^{1/2}\delta_1|).\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{P}\Big[\mathcal{A}_j^C\Big] &= \mathbb{P}\Big[|Z_j| > n^{1/2}|\delta_j|/2\Big] \\ &= 2\overline{\Phi}(n^{1/2}|\delta_j|/2) \\ &\leq 2\overline{\Phi}(n^{1/2}|\delta_1|/2) \\ &\leq 2n^{-1/2}|\delta_1^{-1}|\phi(n^{1/2}|\delta_1|).\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{P}\Big[\mathcal{B}\Big] &= 1-\mathbb{P}\Big[\mathcal{A}_0^C \cup \mathcal{A}_1^C \ldots \mathcal{A}_{p_1}^C\Big] \\ &\geq 1-\sum_{j=0}^{p_1}\mathbb{P}\Big[\mathcal{A}_j^C\Big] \\ &\geq 1-2p_1n^{-1/2}|\delta_1^{-1}|\phi(n^{1/2}|\delta_1|)-4(p-p_1)n^{-1/2}|\delta_1^{-1}|\phi(|n^{1/2}\delta_1|) \\ &\geq 1-(3p_1-4p)n^{-1/2}|\delta_1^{-1}|\phi(|n^{1/2}\delta_1|).\end{aligned}$$

The above calculations imply that $\tilde{U}_j < Cn^{-1/2}|\delta_j^{-1}|\exp\{-n\delta_j^2/2\}$, for some constant $C > 0$ with exponentially high probability.

Define the second difference among ordered p-values as $\Delta_j = P_{(j+1)} + P_{(j-1)} - 2P_{(j)}$. On the set $\mathcal{B}$, $\Delta_j = Cn^{-1/2}|\delta_j^{-1}|\exp\{-n\delta_j^2/2\}$ for $j = 2, \ldots, p_1 - 1$. For

$j = p_1 + 1, \ldots, p$, we have $\mathbb{E}\Delta_j = 0$. Now notice that $\mathbb{E}\Delta_{p_1} = O((p - p_1 + 1)^{-1})$ as well. Consequently, the absolute value of the second derivative peaks near p_1.

□

Theorem 3.5.1 is the main motivation for the methodology that we develop in this paper. If we plot the graph of ordered p-values $P_{(j)}$ (or some monotone non-decreasing transformation thereof) against j/p and imagine a smooth curve through the points, such a curve would have highest curvature (absolute value of second derivative) at p_1/p, which identifies the exact cases for which the null hypothesis may not be true. This feature is also visible in panel (c) of Figure **??**. Naturally, in our algorithms we do not fit a smooth curve first, but explore various ways in which a sharp change in the second derivative may be detected.

Since in many applications where multiple comparisons are considered the dimension p is also large, our next result addresses this case. Correct inference, in the sense of having none of the true null hypotheses rejected (no false positives) and all of the true alternatives not rejected (no false negatives), depends on the *minimal effect size* given by $|\delta_1|$, apart from sample size n and dimension p and to some extent on the number of true alternatives p_1.

Theorem 3.5.2. *Suppose that the dimension $p \equiv p_n$, the number of true alternatives $p_1 \equiv p_{1n}$, and the values of $\delta_j \equiv \delta_{jn}$ are functions of the sample size n, also define the minimal effect size $\delta_n = \min_{j:\mathcal{I}_{H_{1j}}=1} |\delta_{jn}|$. Assume that $p_n \to \infty$ as*

$n \to \infty$. Define the sequences

$$\alpha_{1n} = p_{1n}/p_n,$$
$$\alpha_{0n} = p_n \exp\{-n\delta_n^2/2\}.$$

The probability of correct inference, in which all the true null hypothesis are not rejected while all hypotheses where the null is false are rejected, tends to one if either $\alpha_{0n} \to 0$ or $\alpha_{1n} \to 1$ as $n \to \infty$.

Before providing a sketch of proof of this Theorem, we discuss the implication of this result. Since in many applications of multiple testing the actual number of true alternatives is small relative to the number of tests, the case where $\alpha_{1n} \to 1$ is potentially rare. However, if that is the case, from the proof of Theorem 3.5.1 we realize that most of the p-values would asymptote to zero, while the few cases of true null hypotheses p-values would asymptote to one, thus making correct inference relatively easy.

The more interest case is that of $\alpha_{0n} \to 0$ as $n \to \infty$. Note that while $p_n \to \infty$, this effectively restricts the rate of growth of p_n relative to sample size n. Theoretical results in modern high-dimensional statistics generally require the condition that $n^{-1} \log p_n \to 0$, see for example, multiple results in BÃijhlmann and Geer [4] . We obtain that even under this near-exponential dimension growth condition, it is possible to get asymptotically correct inference, if $\delta = O(1)$. If p_n grows at a polynomial rate of n, then asymptotic correct inference is possible even when $n\delta_n^2 \to \infty$.

Sketch of Proof of Theorem 3.5.2. We build from the proof of Theorem 3.5.1,

where we established that p-values corresponding to the cases where the alternative is true are bounded by $\exp\{-n\delta_n^2/2\}$, while the smallest p-value corresponding to the cases where there the null hypothesis is true has a cumulative distribution function $1-(1-u)^{p_n-p_{1n}}$ for $u \in (0,1)$. Thus the probability that the a correct inference happens is

$$\begin{aligned}
&\left(1-\exp\{-n\delta_n^2/2\}\right)^{(1-\alpha_{1n})p_n} \\
&= \left(1-\alpha_{0n}/p_n\right)^{(1-\alpha_{1n})p_n} \\
&\to \exp\big\{-\alpha_{0n}(1-\alpha_{1n})\big\}.
\end{aligned}$$

This tends to one under the conditions of the Theorem.

□

Our next result is related to one of the geometric methods of multiple testing that we discuss in this paper. We call it the *level-plot* method, where we simply compute the probabilities $\mathbb{P}\big[U_j \leq \alpha_n\big]$ for a specified sequence $\{\alpha_n\}$. In our computations, once α_n is chosen reasonably, this method offers a sharp distinctive plot for elicitation of true alternatives.

Theorem 3.5.3. *Suppose $p_n \to \infty$ and $n^{1/2}\delta_n \to \infty$ as sample size $n \to \infty$, and let $\{\alpha_n\}$ be a sequence of positive reals in $(0,1)$ satisfying*

$$\begin{aligned}
\alpha_n &= O(p_n^{-1}), \text{ and} \\
n^{-1/2}\delta_n^{-1}\alpha_n^{-1}\exp\big\{-n\delta_n^2/2\big\} &= o(1).
\end{aligned}$$

Then

$$\mathbb{P}\big[U_j > \alpha_n\big] \to 0, \ \textit{for } j = 1, \ldots, p_1,$$

$$\mathbb{P}\big[U_j > \alpha_n\big] \to 1, \ \textit{for } j = p_1 + 1, \ldots, p.$$

Moreover, if the decision to reject and not reject the null hypotheses is made based on whether $\mathbb{P}\big[U_j > \alpha_n\big]$ *exceeds* q *or not for some* $q \in (0, 1)$*, then the probability of both type-I an d type-II errors are exponentially small.*

We omit the proof of this Theorem, which follows arguments similar to those used in the proof of Theorem 3.5.1. Note that the U_j's are (functions of) p-values, consequently Theorem 3.5.3 is about probabilities relating to random probabilities, and its practical implementation requires estimation of how often U_j exceeds α_n. We use resampling to compute estimates of $\mathbb{P}\big[U_j \leq \alpha_n\big]$, described in greater detail elsewhere in this paper. Also, for practical purposes, the choice $\alpha_n = 0.05/p$ suffices and this is what we have used for our simulations and data analysis reported below. Note that for Theorem 3.5.3 we *do not* require any kind of control on the rate of growth of p_n with respect to n.

3.6 Simulation study on known p value curve

Several widely used multiple testing methodologies work on vector of p values from the hypothesis tests. Here we consider an experiment, where we generate the ordered p values, according to some known geometry, and try to detect the change in distribution. As a consequence of 3.5.1, we know that left side this change point would reflect the tests where alternative is true.

For generation of the curves, we use the same set of equations as in 3.4.4. We compare the methods that are readily available like Holm, Hochberg, Hommel, Bonferroni, false discovery rate and no correction. In figure 3.4 we show the result of the recovered change points with different comparison methods.

3.7 Real Data Analysis

While there are many common examples of multiple testing in scientific experimental studies, one of the most prominent area requiring multiple testing is Genomics. With the ease of gene sequencing, it is common to see data sets analyzing multiple genomes and producing large data sets. With advances in sequencing technology, it has become a norm to share the data from these experiment publicly with the scientific community. We will consider a microarray data on gene expressions here, however many other areas including m-RNA or protein folding experiments data can also be appropriate here.

The data from many large genome wise association study is publicly available. The common traits for these data sets is that there are a large number of statistical tests are performed over various features. A typical feature can be gene or any other bio marker. For each case, they are tested against a common null hypothesis. Many cases the number of observations for each feature or traits are small compared to the number of tests. Traditional multiple testing methods under these situations tend to produce extremely conservative result.

We consider the data set analyzed by JD [19], Storey and Tibshirani [34].

The data is available at `https://research.nhgri.nih.gov/microarray/NEJM_Supplement/`. The work by Hedenfalk et al Hedenfalk et al. [14] considers BRCA1 and BRCA2 mutation positive tumors using several microarrays from each tumor type. Their goal was to find traits that are significantly different between these two types of cells. They have used a modified F statistic and simultaneously tested each gene to produce p values corresponding to 3226 genes. The box plots for the two types of cells are shown in figure 3.7.

Using 3226 genes, we perform two sample t tests to produce equal number of p values. The distribution of the observed p values are shown in figure 3.6. The original authors have used a p value cut off of 0.001 to select 9-11 genes that show differential expression. However, they have used a different test statistic from what we are using.

Storey et al Storey and Tibshirani [34] have used the method of q values to construct a null distribution, with frequency of p values near 0.67. In their result, using q value of 0.05 they have found 160 genes to be significant with estimated number of false positives close to 8. However, they do acknowledge that the 0.05 cutoff for q values is arbitrary and possibly explain large difference in their result with the original authors.

We performed adjustment of the p values using existing methods including Bonferroni, Hochberg, Holm and FDR. The Bonferroni method is trying to protect against a single false positive and extremely conservative. The figure 3.7 shows the p values after adjustment using Bonferroni method. We used the cutoff of 0.05 to get 2 significant genes.

	Holm	Hochberg	Hommel	Bonferroni	BH	BY	FDR	None
Significant p values	2	2	2	2	44	0	44	598

Table 3.1: The number of significant genes after adjustment by the common methods of adjustments. We have used a cut off of 0.05 to get the significant p values.

The table 3.1 shows the number of significant genes found using 8 adjustment methods. We have used a cutoff of 0.05 in this case to get the significant genes. The method closest to the original result is the FDR producing 44 significant genes.

For our method, we first needed to generate Bootstrap distribution of the p values. We used a Bootstrap sample size of 100 for this. In figure 3.8, we show a comparison of the distribution of the mean of the bootstrapped p values with the original p values. The distribution of the mean of the p values is also shown. We have used the scale $-\log(1-p)$ here on both original and bootstrapped values.

The movements of the point 0.05 as we are changing the cutoff for the hit statistic is shown in figure 3.9. Most of the significant movement is in the region $(0.2, 1)$, where it seems to increase almost in a linear way, which is expected.

Finally, the plot of the quantiles of the order statistics of the bootstrapped p values is shown at figure 3.10. Visually it is quite difficult to determine the position of the hinge from these plots, which is most probably due to very few samples in this data set. However, using the 90th quantile, the cutoff is calculated to be at 40, using both the methods of cumulative sum and linear index regression. This is the result that matches most closely with the original authors findings and also, very close to the result by the FDR method.

3.8 Application to Functional Data

Even though functional data has been gaining popularity as a method of analyzing continuous steams of data in recent years, there has been relatively less work in multiple hypothesis testing in functional data. This is very different from the other applications of multiple testing as it is a widely studied research area in statistics. From our limited literature survey, we came across some work by Pomann et al. [28] which studies two sample tests and cites some other similar works. Most notably, Zhang et al. [43], studies bootstrap based two sample tests for functional data. Some works on functional ANOVA by Cuevas et al. [5] and EstÃl'vez-PÃl'rez and Vilar [10] are also there. Tests on shape of the mean function by HorvÃąth et al. [17] is an application of non parametric statistics, specifically basis representation for multiple testing in this context. Likelihood ratio tests for longitudinal functional data were developed by Staicu et al. [33] also.

Our methodology of change detection using bootstrapped distribution of the p values, does not change in the context of the functional data analysis. However as we are focusing on functional data in this document, we would like to show an application of our methodology in this context. Here we use the data on sea level pressure which is studied in detail in chapter 4 for a different application.

3.8.1 Sea level pressure data

We consider a climate data set on sea level pressure from the northern hemisphere within the longitude $20N$ up words. This data set is used for computing the *arctic*

oscillation index, published by NOAA. We use monthly data, with seasonality removed by subtracting the corresponding monthly means from the data. So, this data can be considered as a spatio temporal data set on a location grid of size 29×144, with each curve is a function of time, with 144 realizations in time. We fitted a simple linear model of $Y_{ij}(t) = \beta_0 + \beta_1 t + \epsilon$, regressing the SLP on time. We then looked for the significance of the slope, which signifies time dependence. Simultaneous testing over all the location makes it a multiple testing problem. In figure 3.11 we show the significant regions selected by FDR and in figure 3.12, we show the significant regions picked up by the p value distribution method. Both methods are using a false positive rate 0.05 on individual tests. For our method, we have used the 90th quantile curve for change detection and used the local linear regression method for estimating the change point. It is clear that the bootstrap based method is much less conservative than FDR. Specifically, we are detecting some regions in the mid Atlantic and in the arctic circle which the FDR method ignores.

Bibliography

[1] F. Abramovich and Y. Benjamini. Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, 1996:351–61.

[2] Y. Benjamini and Y. Hochberg. Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics*, 1997:407–18.

[3] H. Bovenhuis. Spelman rj selective genotyping to detect quantitative trait

loci for multiple traits in outbred populations. *Journal of Dairy Science*, 83 (1):173–80.

[4] Peter BÃijhlmann and Sara A. van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer series in statistics. Springer, Berlin, 2011. ISBN 978-3-642-20192-9 978-3-642-20191-2 978-3-642-26857-1. OCLC: 846468238.

[5] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. An anova test for functional data. *Computational Statistics & Data Analysis*, 47(1):111–122, August 2004. ISSN 01679473. doi: 10.1016/j.csda.2003.10.021. URL `http://linkinghub.elsevier.com/retrieve/pii/S016794730300269X`.

[6] E. I. Drigalenko. Elston rc false discoveries in genome scanning. *Genetics Epidemiology*, 1997:779–84.

[7] S. Dudoit. van der laan mj, pollard ks multiple testing. In *Part I*, pages 220–38. Single-step procedures for control of general Type I error rates. Statistical Applications in Genetics and Molecular Biology 2004; 3(1). Hochberg Y. , Tamhane AC Multiple comparisons procedures. Wiley Finner H., Roters M. Multiple hypotheses testing and expected number of Type I errors. Annals of Statistics 2002; 30, 1987.

[8] S. Dudoit and P. J. Shaffer. Boldrick jc multiple hypothesis testing in microarray experiments. *Statistical Science*, 2003:71–103.

[9] S. P. Ellis, M. D. Underwood, and Arango V. Mixed models and. and multiple comparisons in analysis of human neurochemical maps. *Psychiat Res-Neuroim*, 9:111–19, 2000.

[10] Graciela EstÃl'vez-PÃl'rez and JosÃl' A. Vilar. Functional ANOVA starting from discrete data: an application to air quality data. *Environmental and Ecological Statistics*, 20(3):495–517, September 2013. ISSN 1352-8505, 1573-3009. doi: 10.1007/s10651-012-0231-2. URL `http://link.springer.com/10.1007/s10651-012-0231-2`.

[11] C. R. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, 2004:1035–61.

[12] C. R. Genovese, K. Roeder, and L. Wasserman. False discovery control with p-value weighting. *Biometrika*, 2006:509–24.

[13] E. I. George. The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–08.

[14] Ingrid Hedenfalk, David Duggan, Yidong Chen, Michael Radmacher, Michael Bittner, Richard Simon, Paul Meltzer, Barry Gusterson, Manel Esteller, Mark Raffeld, Zohar Yakhini, Amir Ben-Dor, Edward Dougherty, Juha Kononen, Lukas Bubendorf, Wilfrid Fehrle, Stefania Pittaluga, Sofia Gruvberger, Niklas Loman, Oskar Johannsson, HÃěkan Olsson, Benjamin Wilfond, Guido Sauter, Olli-P. Kallioniemi, ÃĚke Borg, and Jeffrey Trent. Gene-Expression Profiles in Hereditary Breast Cancer. *New England*

Journal of Medicine, 344(8):539–548, February 2001. ISSN 0028-4793. doi: 10.1056/NEJM200102223440801. URL http://dx.doi.org/10.1056/NEJM200102223440801.

[15] D. W. Heyen, J. I. Weller, and M. Ron. A genome scan for qtl influencing milk production and health traits in dairy cattle. *Physiological Genomics;*, 1 (3):165–75, 1999.

[16] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 1979:65–70.

[17] Lajos HorvÃąth, Piotr Kokoszka, and Ron Reeder. Estimation of the mean of functional time series and a two-sample problem: *Estimation of the Mean of Functional Time Series. Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):103–122, January 2013. ISSN 13697412. doi: 10.1111/j.1467-9868.2012.01032.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2012.01032.x.

[18] E. H. Ip. Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, 66(1):109–32.

[19] Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*, 31, 2003. bibtex: stbayes.

[20] Storey JD. The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, 31, 2003.

[21] P. Khatri and M. Babyyak. Croughwell nd temperature during coronary artery bypass surgery affects quality of life. *Annals of Thoracical Surgery;*, 71(1):110–16, 2001.

[22] Clive Loader. *Local regression and likelihood*. Springer Science & Business Media, 2006. URL `http://books.google.com/books?hl=en&lr=&id=NpjeBwAAQBAJ&oi=fnd&pg=PA1&dq=%22and+asymptotic+theory.+Largely,+these+chapters+are+independent+of%22+%22is+necessarily+selective.+I+attempt+to+present+results+that+are%22+%22more+general+approach+to+achieve+the+same%22+&ots=wVafYZoO54&sig=ay4Ekmj0zBkQ2hRgLL-kPojqhRQ`.

[23] B. R. Logan. Rowe db an evaluation of thresholding techniques in fmri analysis. *Neuroimage*, 2004:95–108.

[24] E. P. Merriam and C. R. Genovese. Colby cl spatial updating in human parietal cortex. *Neuron*, 2003:361–73.

[25] M. O. Mosig, E. Lipkin, G. Khutoreskaya, E. Tchourzyna, and M. Soller. Fridmann aa whole genome scan for quantitative trait loci affecting milk protein percentage in israeli-holstein cattle, by means of selective milk dna pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics*, 2001:1683–98.

[26] K. J. Ottenbacher. Quantitative evaluation of multiplicity in epidemiology and public health research. *American Journal of Epidemiology*, 1998:615–19.

[27] F. Pesarin. Multivariate permutation tests with applications to biostatistics. In *Wiley*, pages 25–9. 2001. Efron B., Tibshirani R. An introduction to the Bootstrap. Springer-Verlag Troendle K, McShane. An example of slow convergence of the Bootstrap in high dimensions. American Statistician 2004; 58, 1993.

[28] Gina-Maria Pomann, Ana-Maria Staicu, and Sujit Ghosh. A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(3):395–414, April 2016. ISSN 00359254. doi: 10.1111/rssc.12130. URL `http://doi.wiley.com/10.1111/rssc.12130`.

[29] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics;*, 19: 368–75, 2003.

[30] C. M. Schaffer. Green pe cluster-based market segmentation: some further comparisons of alternative approaches. *Journal of Market Research Sociology*, 1998:155–63.

[31] M. Schlaeppi and K. Edwards. Fuller rw patient perception of the diskus inhaler: a comparison with the turbuhaler inhaler. *British Journal of Clinical Practice*, 1996:14–19.

[32] P. Sebastiani, E. Gussoni, and I. S. Kohane. Ramoni mf statistical challenges in functional genomics. *Statistical Science*, 2003:33–70.

[33] Ana-Maria Staicu, Yingxing Li, Ciprian M. Crainiceanu, and David Ruppert. Likelihood Ratio Tests for Dependent Data with Applications to Longitudinal and Functional Data Analysis: Testing for functional processes. *Scandinavian Journal of Statistics*, 41(4):932–949, December 2014. ISSN 03036898. doi: 10.1111/sjos.12075. URL `http://doi.wiley.com/10.1111/sjos.12075`.

[34] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003. URL `http://www.pnas.org/content/100/16/9440.short`.

[35] Dudoit S. van der Laan MJ. Pollard ks augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

[36] K. Vedantham, A. Brunet, and Boyer R. Post traumatic stress disorder. trauma exposure, and the current health of canadian bus drivers. *Canadian Journal of Psychiatrics*, 46(2):149–55.

[37] J. I. Weller and J. Z. Song. Heyen dw a new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics;*, 150(4):1699–706, 1998.

[38] P. H. Westfall. Young ss resampling-based multiple testing: examples and methods for p-value adjustment. In *Wiley*, pages 344–49. 1005-10, Exploring Data Tables, Trends, and Shapes. Wiley, 1985. Ahmed SW Issues arising in

the application of Bonferroni procedures in federal surveys . In 1991 ASA Proceedings of the Survey Research Methods Section . Wright SP Adjusted p-values for simultane ous inference. Biometrics 1992; 48, 1991.

[39] P. H. Westfall, S. Kropf, and L. Finos. Weighted fwe-controlling methods in high-dimensional situations. In Y. Benjamini, F. Bretz, and S. Sarkar, editors, *eds*, pages 143–54. Recent developments in multiple comparison procedures. vol. 47. Institute of Mathematical Statistics Lecture Notes-Monograph Series. Benjamini Y. , Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society (Ser B) 1995; 57: 289-300, 2004.

[40] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, and K. J. Friston. Evans ac a unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73, 1996.

[41] D. Yekutieli, A. Reiner-Benaim, Y. Benjamini, G. I. Elmer, N. Kafkafi, and N. E. Letwin. Lee nh approaches to multiplicity issues in complex research in microarray analysis. *Statistica Neerlandica*, 2006:414–37.

[42] Yosef Hochberg Yoav Benjamini. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL `http://www.jstor.org/stable/2346101`.

[43] Chongqi Zhang, Heng Peng, and Jin-Ting Zhang. Two Samples Tests

for Functional Data. *Communications in Statistics - Theory and Methods*, 39(4):559–578, February 2010. ISSN 0361-0926, 1532-415X. doi: 10.1080/03610920902755839. URL http://www.tandfonline.com/doi/abs/10.1080/03610920902755839.

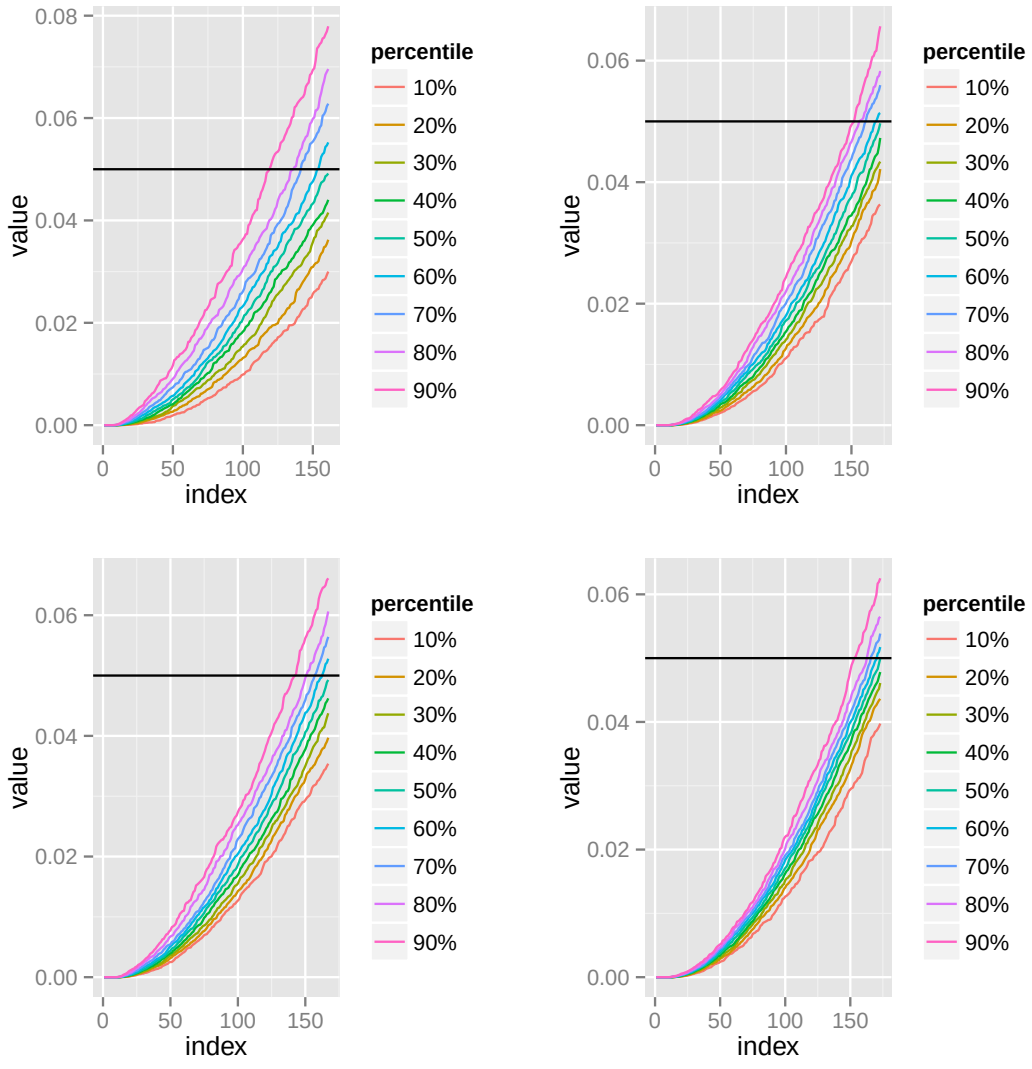


Figure 3.1: Bootstraped edf of p values. Sample size increases from $20, 50, 100, 200$ column wise

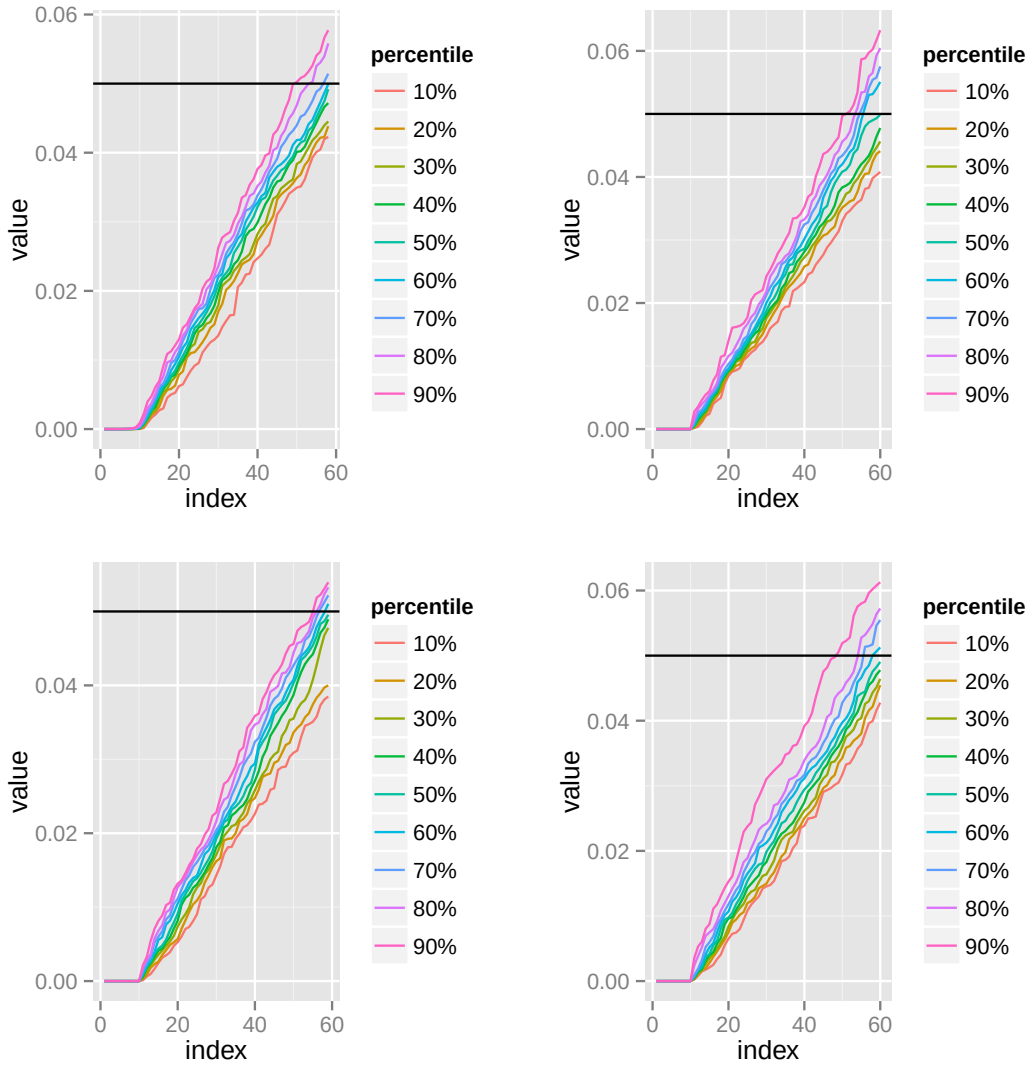


Figure 3.2: independely generated quantiles from edf of p values. Sample size increases from $20, 50, 100, 200$ column wise

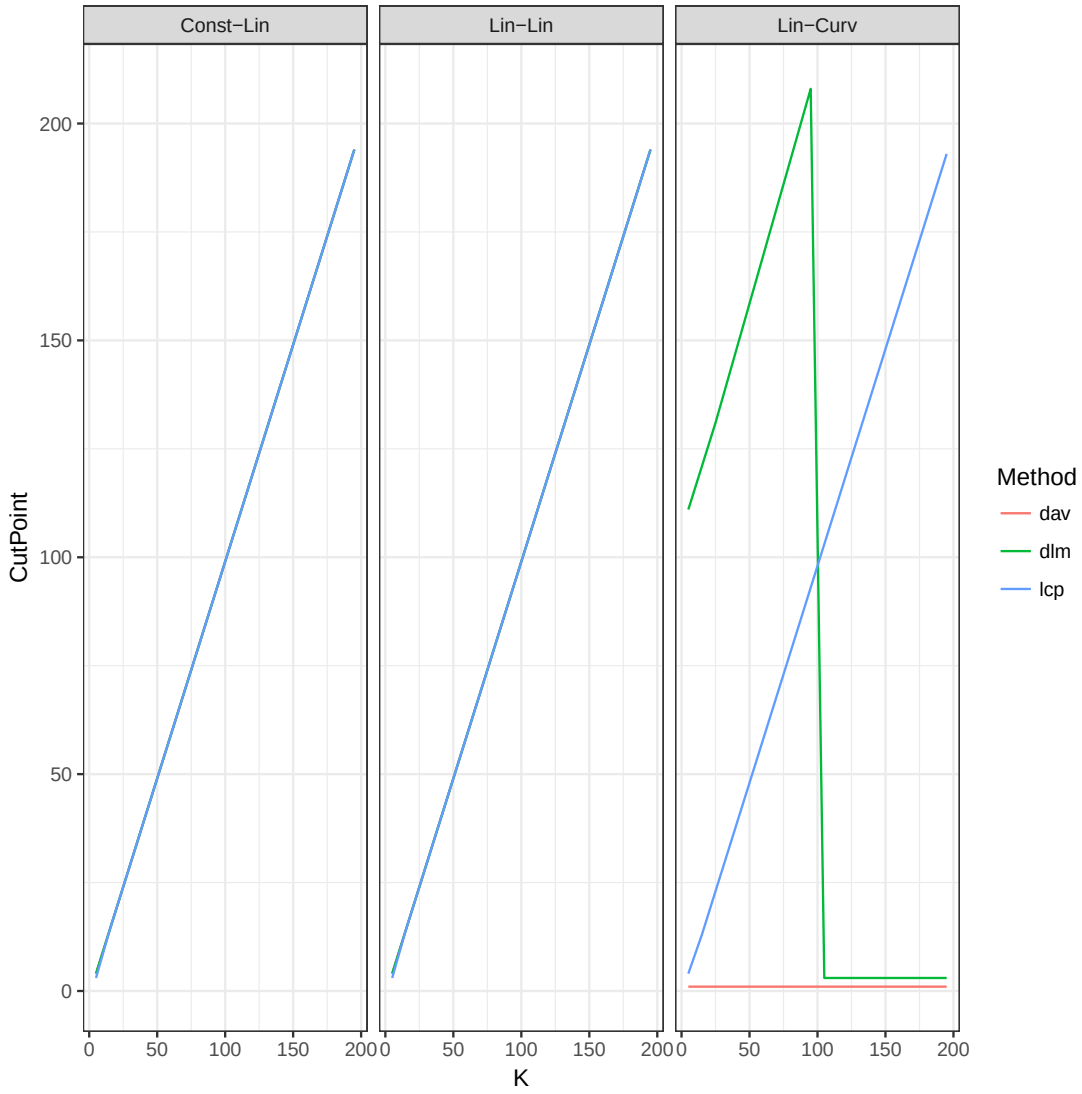


Figure 3.3: Comparison of detection of change points in various scenarios by three methods.

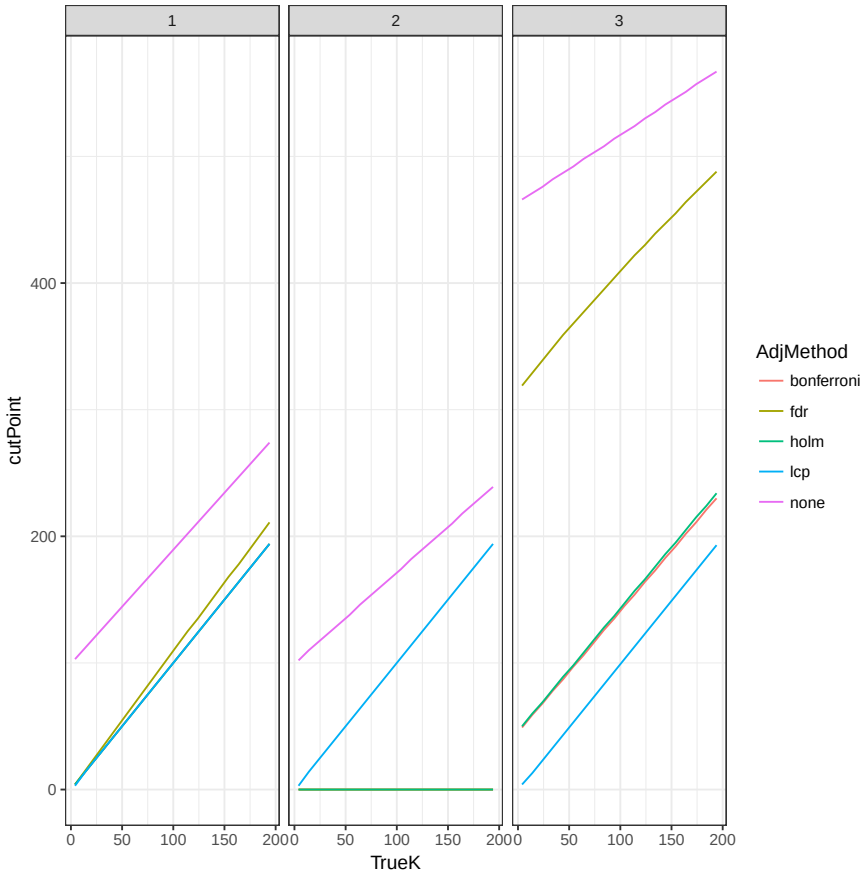


Figure 3.4: Comparison of change detection in distribution of p values using several multiple comparison methods. Three panels indicate different scenarios of distribution of p values specified in 3.4.4. Method 'lcp' indicates bootstrap based method with local linear regression as change detection.

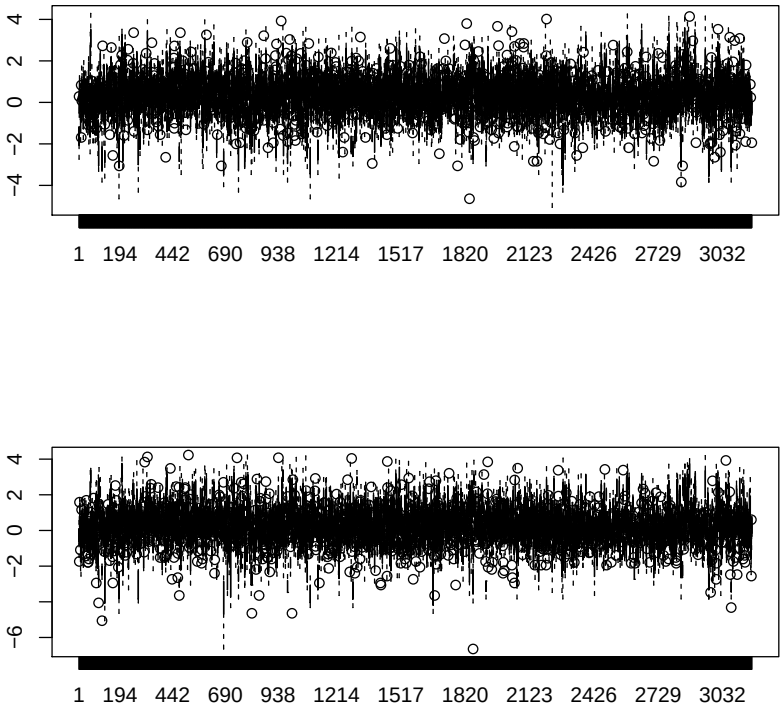


Figure 3.5: Box plots of gene expression values for BRCA 1 (top) and BRCA 2 (bottom) type tumor cells.

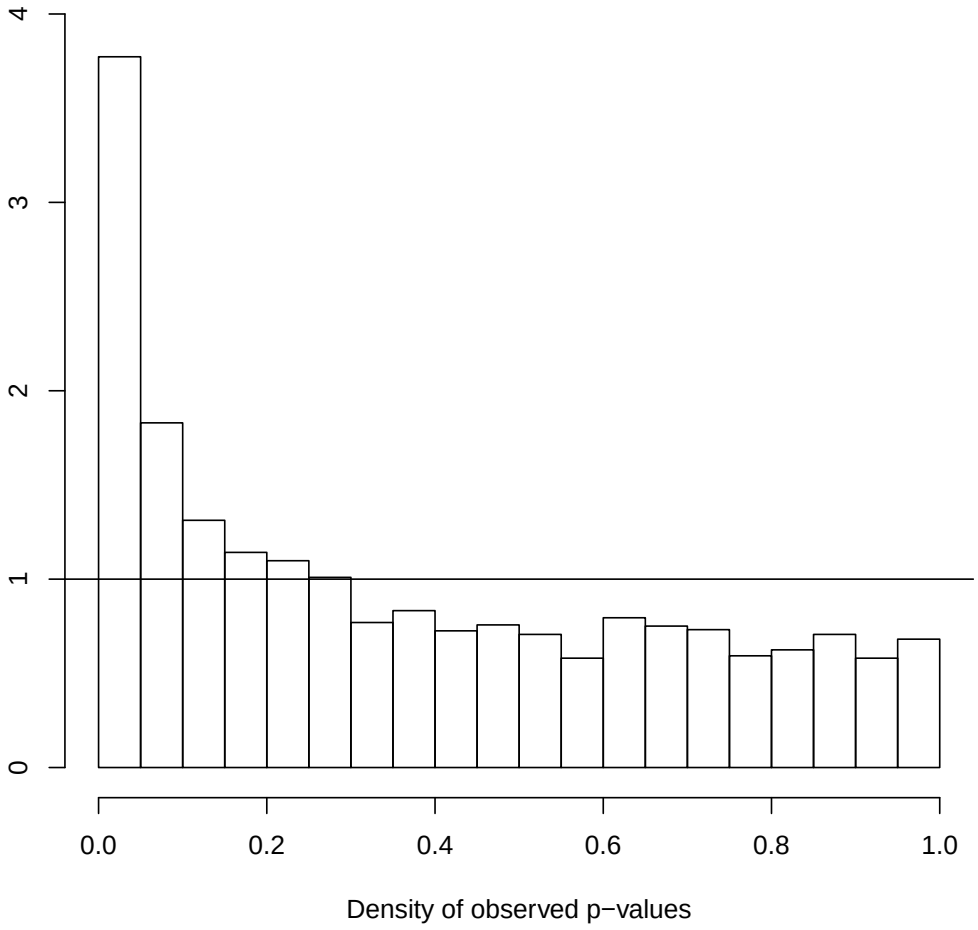


Figure 3.6: Histogram of the observed p values from the BRCA data. The horizontal line shows the profile in case no gene exhibited any differential expression or the case for all nulls are true.

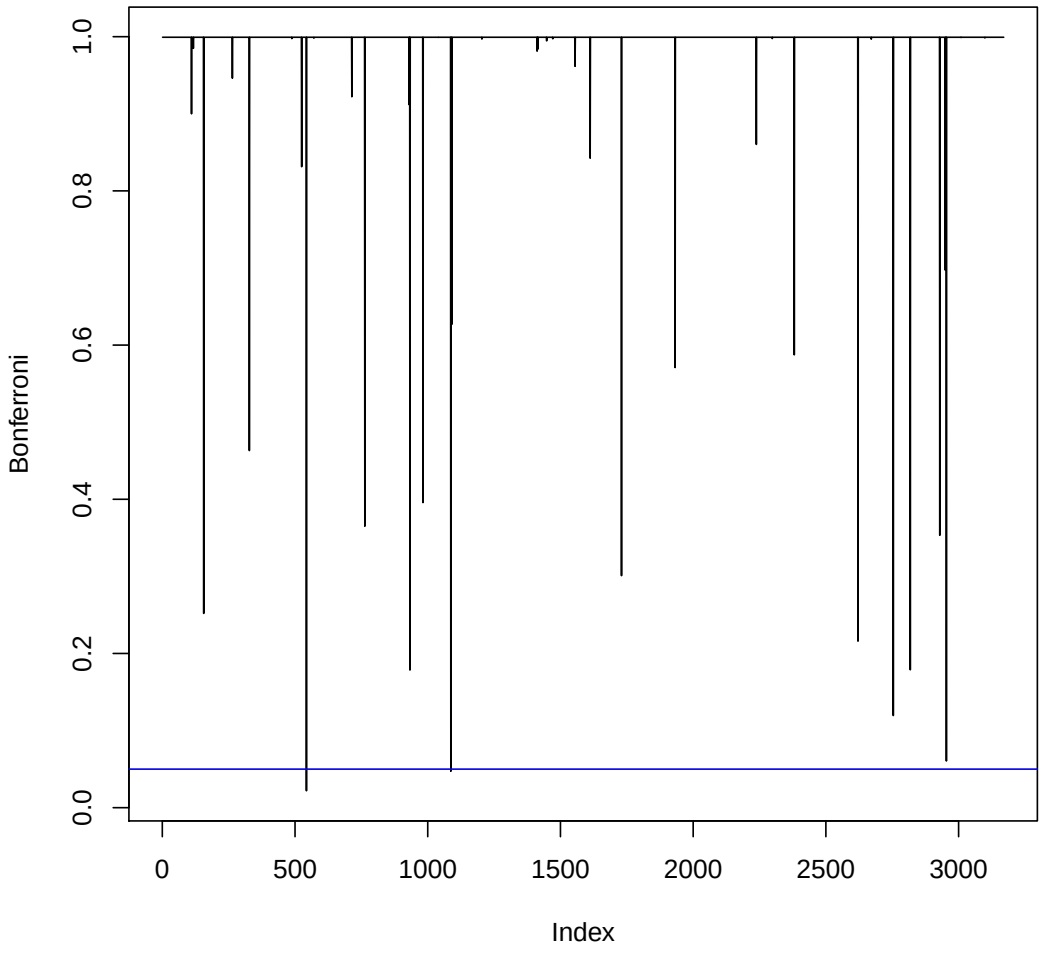


Figure 3.7: Plot of the p values using Bonferroni adjustment. The blue line shows the 0.05 cutoff used.

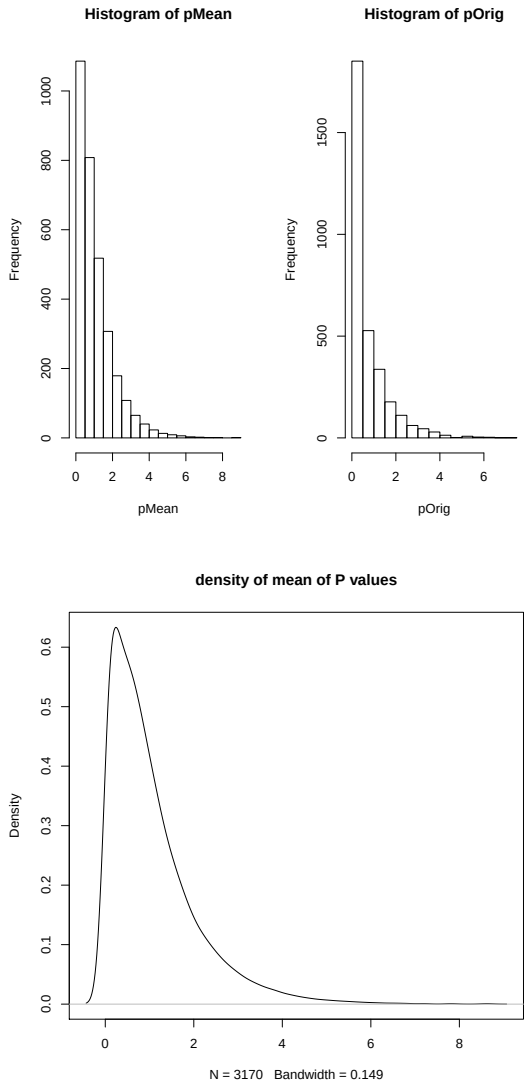


Figure 3.8: (a) Comparison of histograms of Bootstrapped p values and ovserved p values. The Left image shows the mean of the bootstrapped p values and the right image shows the original p values. We are using the transformation $-log(1-p)$ to scale the p values. (b) Density of the mean of the boostrapped p values.

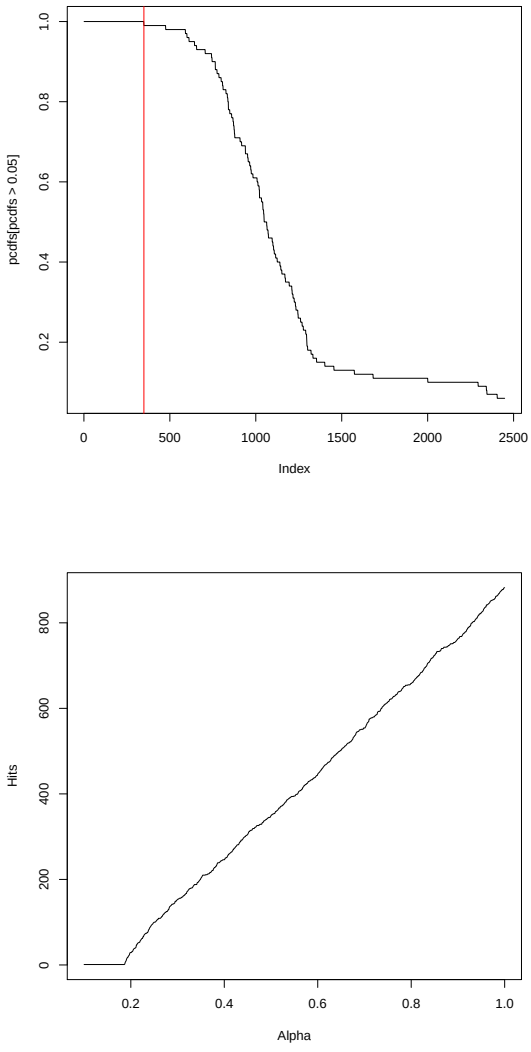


Figure 3.9: (a) Hitplots of Bootstrapped p values with a cutoff at 0.5. (b) The cutoffs with the hit points for all the hit plots with the cutoff varied from 0 to 1

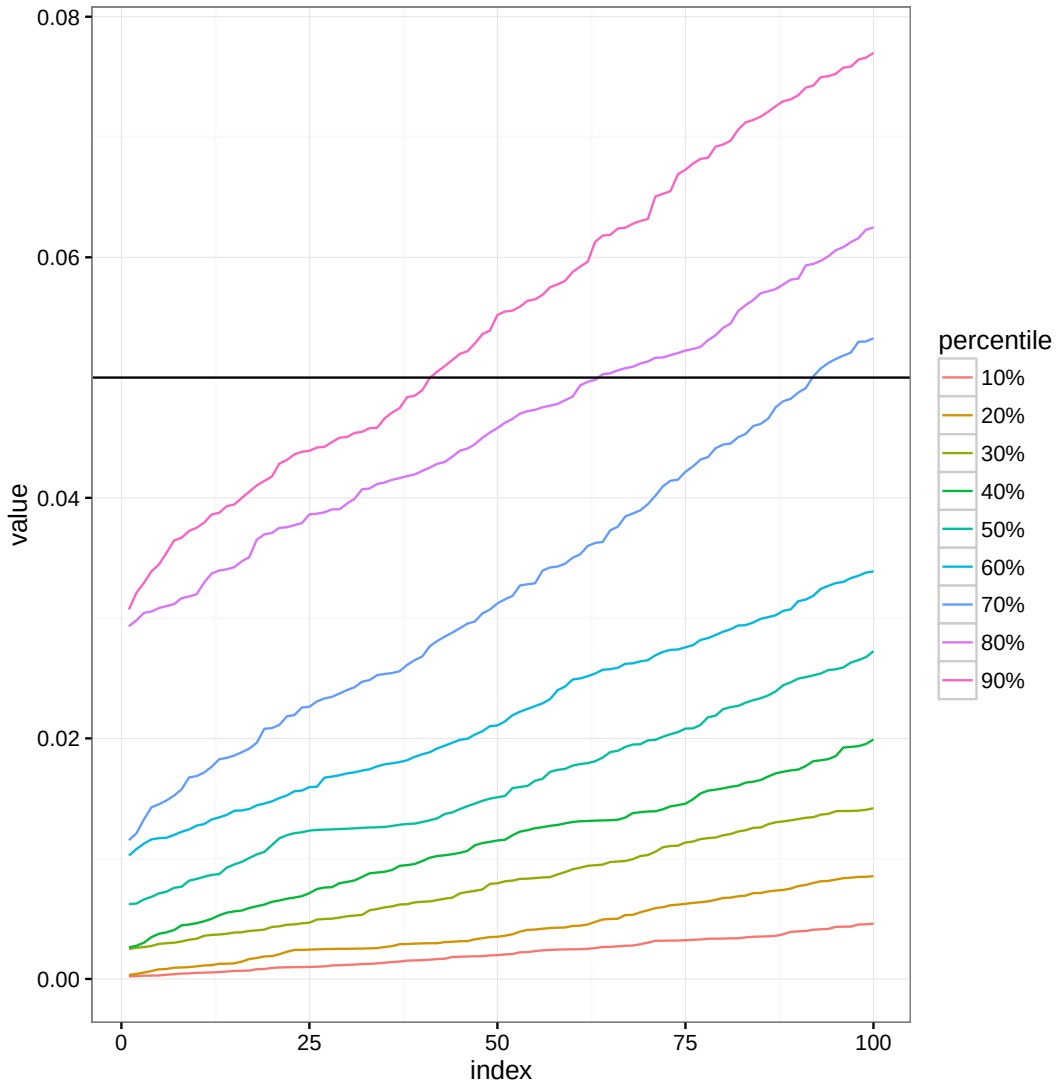


Figure 3.10: Plots of the quantiles of the empirical density of the order statistics of the p values using Bootstrap.

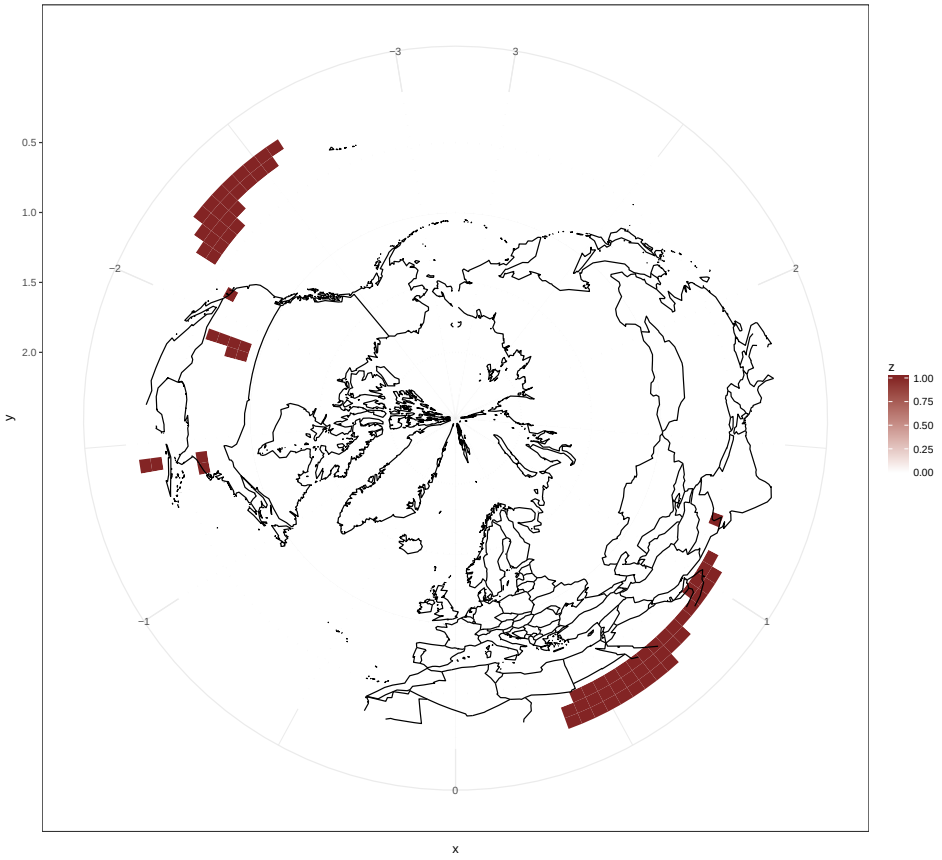


Figure 3.11: Significant regions of Sea level pressure with time using FDR.

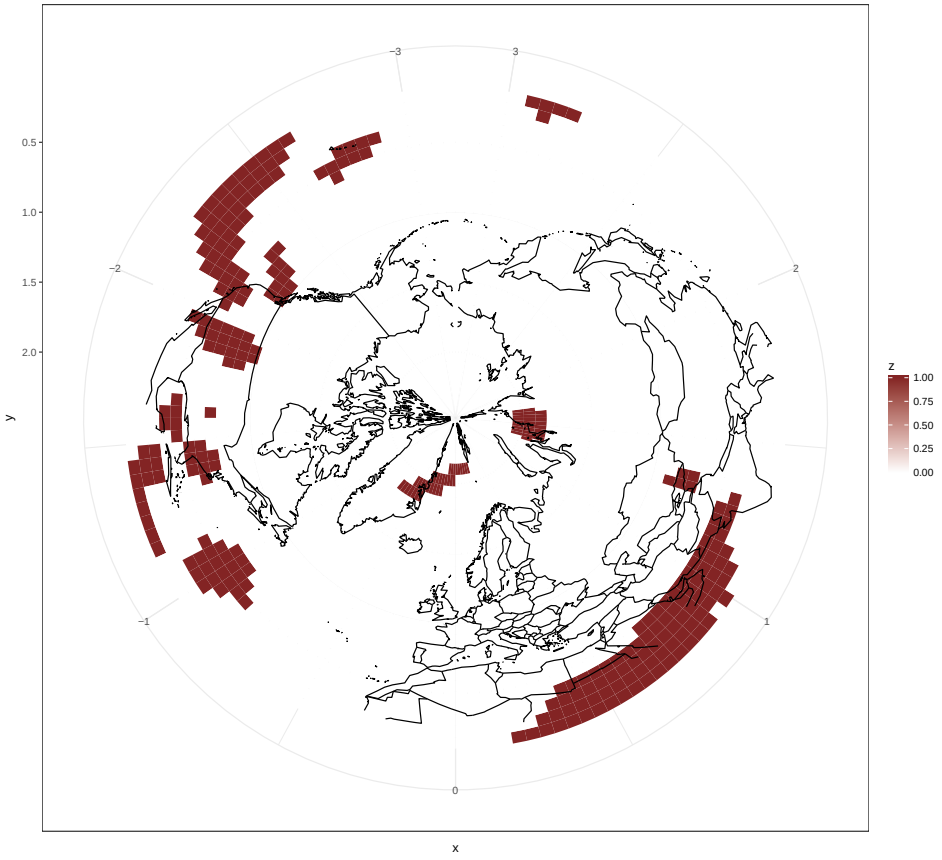


Figure 3.12: Significant regions of Sea level pressure with time using distribution of p values.

Chapter 4

Functional Data Analysis using Envelope Semimetric

4.1 Introduction

We consider the problem of supervised learning from functional random variables with numeric or categorical responses. Let $(\Omega, \mathbb{F}, \mathbb{P})$ be a probability space and $L_2(\Omega, \mathbb{P})$ be the set of all random variables in Ω which are square integrable. For a compact set T, we consider functional random variable $\mathcal{X} : \Omega \times T \mapsto \mathbb{R}$ so that for each $t \in T$, $X(., t)$ is square integrable. We consider the class of supervised learning problems with data in the form of $(\mathcal{X}, Y)$, where $\mathcal{X}$ is a functional valued random variable and Y can be either numeric or categorical. This type of problem is often referred to as functional regression or classification respectively.

With the advent of new and cheap data collection methods and storage options, functional data is becoming common. We refer to Wang et al. [50], Rice [45], Zhao et al. [54], MÃijller and StadtmÃijller [37] for updated review of interesting application of functional data analysis. Several books and monographs are available in this area with plenty of application examples in several related fields, bos [1], Ramsay et al. [43], Ferraty and Vieu [19], Wu and Zhang [52], Hsing and Eubank [27].

The will use a non parametric model to represent this problem. Consider $g : \mathcal{X} \mapsto \mathbb{R}$ a function mapping the functional data into real line, then we can represent the functional regression problem as $y = g(\mathcal{X}) + \epsilon$. Here ϵ are the noise, that we assume to be coming from a normal distribution with constant variance. In the case of categorical response, the function $g(.)$ maps the functional space to categorical space likewise.

Our main objective to fit this model is to estimate the function $g(.)$. In the non

parametric methods literature this problem is handled using a kernel method. A kernel is a distance metric that is used to represent distances in high dimension into fewer dimension, into one dimension in our case. We use a standard Nadaraya Watson kernel in one dimension. However, in our methodology, the kernel is readily replaceable to any other standard kernel.

The problems in functional data analysis is exactly analogous to the multivariate counterpart. We can use the standard notation for supervised learning here. Given we observe $(\mathcal{X}_i, \mathcal{Y}_i)$, where $\mathcal{X}_i$ are the covariates and $\mathcal{Y}_i$ being the response, the problem of supervised learning here is to predict the response for a new observation $\mathcal{X}$. Here, in the case of functional data, usually X_is are functional random variables and Y_is can be either real continuous or categorical. Depending on the nature of Y_i, the problem is usually called classification or regression. While it is possible for Y_i to be functional response, we will not be considering such problems in this document. In case the response Y_is are missing, the problem becomes of functional clustering or unsupervised learning.

This chapter is is structured as follows: we first discuss established techniques in modeling functional data with semi metrics. We then give background of the envelope methodology we will be using. We propose our semi metric with envelopes and explore its theoretical properties. We discuss the estimation method and computation related to to envelope estimation. We then show the benefit of our method using simulation setup. We finally show an application of our method in analysis of Arctic Oscillation data with Sea Ice Concentration pattern.

4.2 Semi metrics in Functional Data

Because the functional random variable takes values in an infinite dimensional space, it is common to conceptualize X taking values in an inner produce space like Hilbert space. This kind of assumption is good for theoretical study of functional data, however with real data applications, we need some approximation of the norms. Following Ferraty [20], we adopt the notion of semi metric space for functional data. A semi metric $d : \mathbb{F} \mapsto \mathbb{R}$ behaves similar to a metric, except $d(x) = 0$ does not imply $x = 0$. The benefit of using semi metric to describe functional data is they are easily interpreted and often easy to implement. Theoretical properties with the semi metric approach has also been studied in the literature.

There has been many semi metrics proposed in the context of learning problems with functional valued predictors. Functional principal component analysis is analogous to PCA in multivariate case for the functional data. The development of FPCA has a long history in this area, see Dauxois et al. [17], Castro et al. [9], Shi [47], Locantore et al. [32] for some of the early work in this method. For a detailed description of theory and applications of FPCA we refer to Jolliffe [30]. FPCA for fully and densely observed functional data was studied in Besse and Ramsay [4], Boente and Fraiman [5], bos [1], Hall and Hosseini-Nasab [24], Pezzulli and Silverman [40], Cardot [6] and many other prominent works in this field.

Most common example of a semi metric in the context of functional data is functional principal component distance. For square integrable functional random variables, the covariance operator can be decomposed into it's eigen functions. PC distance is approximated using first few components of the Eigen functions. This

idea of decomposing a functional variable into orthonormal components is also used for basis expansion methods like Fourier or B-Spline basis decomposition. These methods gives rise to the basis metrics (Fourier or Spline) for functional data. Some semi metrics can be constructed using the parametric structure of the basis functions as well. In case of Fourier or some other basis, the derivatives contain important information. Hence the weighted sum of the derivatives are also used as semi metrics. Partial least squares regression, which is popular in the fields of chemometrics and other applied sciences has also been used to construct semi metrics for functional data Preda and Saporta [42]. We refer to Ferraty [20] for definitions of the many of the semi metrics used in the literature.

The choice of proper semi metric is very important because the topological properties defined by the metric would control the concentration properties of the functional space. Most of the asymptotic results in the literature depends on the concentration properties, hence a suitable metric can improve the efficiency of the learning algorithms greatly.

With a semi metric of choice, we use kernel method to fit our model. Kernels are non parametric method, used to locally weight observations. It is a very popular method for non parametric statistics for its simplicity and ease of application. A kernel function is a weighting scheme that can be used to smooth the observations by averaging them locally. Most kernels would have a bandwidth parameter that controls the degree of smoothness that can be achieved. The estimation of the bandwidth parameter is very important in this type of methodology as there is a bias variance trade off existing here.

A large number of kernel choice is available in the literature, we refer to HorovÃą et al. [26] for a collection of newer advances. In the case of multivariate data, kernel function is generally applied to a norm (semi metric, in our case). We will consider the Gaussian kernel for its smoothness properties.

In the following sections, we will briefly review the developments with envelope methods, which is crucial for out methodology. We will then define our semi metric and show its application in the functional data analysis case.

4.3 Mean Envelopes

The idea of envelopes, introduced by Cook et al. [16] was developed as a dimension reduction tool for multivariate linear model. Envelopes are a class of methods improving efficiency of the coefficient estimates for the multivariate regression. It achieves better efficiency through decomposing the variation into *material and immaterial* components in such a way that the material part aligns with the estimate of the coefficient. This kind of targeted dimension reduction is especially efficient when the immaterial variation outweighs the material component.

Even though originally developed for linear model context, envelopes are a general methodology and we will look into the application on estimation of multivariate mean. More specifically, if $X_1, \cdots, X_n$ are iid p variate normal $N(\mu, \Sigma)$, we consider the application of envelopes in estimating μ. One simple estimator of μ in this case is sample mean $\bar{X}$, however, this does not contain any information from the variability of the data. We try to decompose the mean into $\mu = \Gamma\eta$, such

that Γ is a reducing subspace of Σ. To explain the implications of projecting μ into such space is explained through the following lemma.

The envelope estimator would reside in a subspace $\mathbb{R}^u \subseteq \mathbb{R}^p$, such that the projection $P_S = \Gamma\Gamma^T$ of the data into S is independent of the projection into the orthogonal component $Q_S = I_p - P_S$. It is to be noted that P_S is a symmetric matrix of dimension $p \times p$.

$$P_S X \perp Q_S X \text{ and } \mu \in S \tag{4.1}$$

Existence of such a space trivially holds true as the above condition holds for the full space $\mathbb{R}^p$. However, we are interested in the cases where S is of smaller dimension.

In their introductory paper Cook et al. [16] considers envelopes in a larger generalization. However, in a simple application as ours, we need to explicitly state the motivation behind construction of envelopes in the following lemma 8.

Lemma 8. *The condition (4.1) is true if and only if*

$$\Sigma = (P_S \Sigma P_S + Q_S \Sigma Q_S) \tag{4.2}$$

Proof. As $X \sim N(\mu, \Sigma)$, pre multiplication of X by P_S would be $P_S X \sim N(P_S\mu, P_S\Sigma P_S^T)$. As the projection matrix P_S is symmetric, $P_S X \sim N(P_S\mu, P_S\Sigma P_S)$. As $\mu \in S$, $Q_S X \sim N(0, Q_S\Sigma Q_S)$. For multivariate normal, $P_S X \perp Q_S X$ holds if and only if $P_S\Sigma Q_S = 0$. Hence, $\Sigma = (P_S + Q_S)\Sigma(P_S + Q_S) = P_S\Sigma P_S + Q_S\Sigma Q_S$ □

So we see that the covariance matrix is decomposed into two parts, which are controlled by the projection to the envelope subspace S and its orthogonal

direction. Thus, the envelope projection implies decomposition of the variation towards the mean and its orthogonal direction. These two parts are referred to as material and immaterial information in the envelope literature.

The semi orthogonal basis for the envelope space is $\Gamma \in \mathbb{R}^{p\times u}$. We can reparametrize the covariance matrix as well following 8. We recall that the projection is defined as $P_S = \Gamma\Gamma^T$ and its orthogonal component is $Q_s = I_p - P_s \equiv \Gamma_0\Gamma_0^T$.

$$\begin{aligned}\Sigma =& P_S\Sigma P_S + Q_S\Sigma Q_S \\ =& \Gamma\Gamma^T\Sigma\Gamma\Gamma^T + \Gamma_0\Gamma_0^T\Sigma\Gamma_0\Gamma_0^T \\ =& \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T\end{aligned} \tag{4.3}$$

So, combing both results, the data is parametrized as

$$X \sim N(\Gamma\eta, \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T) \tag{4.4}$$

4.4 Envelope based semi metric

We can now use the concept of envelope into defining a semi metric for functional data. Let $X(t)$ be a functional random variable and we are observing the realization at time points $t = \{1, \cdots, T\}$. We assume these are equispaced intervals and small enough for the curve $X(t)$ to be sufficiently smooth. So the observed data set $\{X_1(t), \cdots X_n(t)\}$ can be assumed to be of a T dimensional normal distribution. Let $\Gamma \in \mathbb{R}^{T\times u}$ be the orthonormal basis for the envelope space in this

context. We denote the elements of the basis as T dimensional vectors representing as $\Gamma = \{\gamma_1, \cdots, \gamma_u\}$. Then the envelope semi-metric is empirically defined as

$$d_u(X_i, X_j) = \sqrt{\sum_{k=1}^{u}(\sum_{t=1}^{T}(X_i(t) - X_j(t))[\gamma_k(t)])^2} \tag{4.5}$$

Here the envelope subspace is the reduced subspace of the covariance matrix Σ. We can also use a weighting matrix to control the smoothing in the above distance. In that case $\gamma_k(t)$ would be basis vectors of the weighted covariance matrix ΣW, where $W = diag(w_1, \cdots, w_T)$ defines the weights on each time point. In that case the distance would be defined as

$$d_u^w(X_i, X_j) = \sqrt{\sum_{k=1}^{u} w_t(\sum_{t=1}^{T}(X_i(t) - X_j(t))[\gamma_k(t)])^2} \tag{4.6}$$

4.5 Application in Functional Data Analysis

Our application of the semi metric in learning problems on functional data will follow the approach by Ferraty [20], which is non parametric and kernel based. A kernel function $K(.)$ is a local weighting function that is used to transform the observed instances of a random variables into their distance weighted versions. For example, a typical weighting would be transforming $X_1, \cdots, X_n$ into $\Delta_1, \cdots, \Delta_n$, where,

$$\Delta_i(x, h, K, d) = \frac{1}{h}K(\frac{d(x, X_i)}{h}) \tag{4.7}$$

Here $d(x, .)$ denote the distance from the fixed x in semi metric $d(.)$. The

bandwidth parameter h is used to control the size of the weighting function. The kernel function can be symmetric or asymmetric, having only positive values. There is a vast literature on types of kernel and the choice of the bandwidth parameter. This is a common approach in non parametric statistics.

We propose to use the envelope distance defined in 4.6 in the local weighting function 4.7. Kernel based methods can be applied on both functional classification and regression. We briefly recall the two major types of learning problems in functional data.

If we observe $(X_i, Y_i) \in \mathbb{F} \times \mathbb{R}$ for $i = 1, \cdots, n$, where X_i are iid functional random variables and Y_i are real valued response. Given a new functional observation X we would like to predict y, which is the problem of functional regression. A simple prediction would be to use the conditional expectation of the response on the predictor.

$$\begin{aligned} \hat{Y}(X) =& E(\mathbb{Y}|\mathbb{X} = X) \\ =& \frac{\sum_{i=1}^{n} Y_i K(h^{-1} d_u^w(\mathbb{X}, X_i))}{\sum_{i=1}^{n} K(h^{-1} d_u^w(\mathbb{X}, X_i))} \\ =& \sum_{i=1}^{n} \omega_{i,h,u}(\mathbb{X}) Y_i \end{aligned} \tag{4.8}$$

Here we have denoted the kernel weights by $\omega_{i,h,u}$ as they depend on the bandwidth and the dimension of the envelope subspace. $\sum_{i=1}^{n} \omega_{i,h,u} = 1$ obviously for all (h, u).

If the response $Y_i \in \{1, \cdots, G\}$ is categorical, then the problem becomes a functional classification problem. In case Y_i is not present, then this would be

unsupervised and is functional clustering. The probability for a class can be calculated similarly.

$$\begin{aligned}\hat{p}_g(\mathbb{X}) =& EI(Y_i = g|\mathbb{X} = X) \\ =& \frac{\sum_{i=1}^{n} I(Y_i = g)K(h^{-1}d_u^w(\mathbb{X}, X_i))}{\sum_{i=1}^{n} K(h^{-1}d_u^w(\mathbb{X}, X_i))} \\ =& \sum_{i:Y_i=g}^{n} \omega_{i,h,u}(\mathbb{X})\end{aligned} \tag{4.9}$$

The definition of the kernel weights $\omega_{i,h,u}$ are same as in the regression case. Both of these methods generalize any nearest neighbor based methods as well.

In the literature many forms of functional regression has been studied. Our problem of interest is vector responses (can be univariate) with functional predictor. The other types of problems considered are functional responses with functional predictors, functional responses with vector predictors. However, the problem of scalar response with functional predictors has been most widely studied in the literature. For a review of methods and latest developments in this area we refer to MÃIJLLER [36], Morris [35].

Most common models for functional predictors with scalar response is functional linear model, where the response is modeled as a linear function of the inner product of the functional covariate and a coefficient vector. This kind of model has been reviewed in Hall and Horowitz [23], Cardot et al. [8], Cardot et al. [7]. Any family of orthonomal basis functions can be used to expand the covariates and to estimate the coefficient vectors. Fourier or B-spline are popular choices of basis functions with the number of coefficients determined usually through cross

validation.

Our approach is different from the functional linear model because we do not assume linear relationship among the predictors and response. We use a kernel function to approximate the non linear relationship among them. There has been several attempts at extending functional linear model with a link function to model non linear behavior. We refer to James [29], Cardot et al. [8], Wang et al. [51] and Dou et al. [18] for a comprehensive review of the generalized functional linear models. There has been some work when the link function is unknown as well, we refer to MÃijller and StadtmÃijller [37], Chen et al. [12] for interesting work on single index models. Functional linear models with unknown variance function has also been explored and referred to as multiple functional index models. The works in this approach of modeling is related to sliced inverse regression and sufficient dimension reduction methods.

The functional regression problem we are working with, can be written as

$$Y = \mu(\Gamma X) + \epsilon \tag{4.10}$$

Here Γ is the estimated basis for the envelope, so we are modeling the projection of the data into the envelope subspace. Our estimate for the mean function would be denoted by $\hat{\mu}_{K,h}$. The kernel weighted smoothing estimator is used to define the function $\mu_{K,h}(.)$. The kernel choice $K(.)$ and the bandwidth choice h used to define the kernel function determines this function.

In general, the two basic problem of supervised learning from functional data can be summarized depending on the type of response variable. If we denote the (response, predictor) pair by $(Y, X(t))$, where $X(t)$ is functional with index

variable t. Depending on if Y is categorical or numeric, with the observed version of $X(t)$, we can write the relationship between X and Y in 4.11.

$$\left\{ \begin{array}{ll} Y = \xi_{k,h}(\Gamma_u X), & \text{Y categorical} \\ Y = \mu_{k,h}(\Gamma_u X), & \text{Y numerical} \end{array} \right\} \tag{4.11}$$

The functions $\xi_{k,h} : \mathcal{X} \mapsto \mathbb{L}$ and $\mu_{k,h} : \mathcal{X} \mapsto \mathbb{R}$ are unknown relationships between the predictor and response that we will estimate. Here $\mathbb{L}$ denote the space of possible labels the response Y can accept when it is categorical. The subscripts (k, h, u) denote the choice of kernel, bandwidth and dimension of the envelope subspace respectively.

4.6 Quantile Estimators

The estimator given in 4.25 can be considered as an weighted mean of the responses, with the weights defined by the kernel function given as

$$W_i(\mathcal{X}) = \frac{K(h^{-1} d_{u,\Gamma}(\mathcal{X}, \mathcal{X}'))}{\sum_{i=1}^{n} K(h^{-1} d_{u,\Gamma}(\mathcal{X}, \mathcal{X}'))} \tag{4.12}$$

However, in many cases, using weighted average is not the best option, especially when the responses are skewed or have a heavy tailed distribution in case of continuous responses. In those cases, it might be better to use quantile based estimators.

To define the conditional quantile, we first need to specify the conditional distribution $\mathcal{F}_{Y|\mathcal{X}}$. The conditional CDF is defined as

$$\mathcal{F}_{Y|\mathcal{X}} = P(y \leq Y|\mathcal{X}) \tag{4.13}$$

It is assumed that such a conditional distribution exists and the conditions to ensure the existence of this distribution is rather complicated and left out of the scope of present discussion. Now, it is easier to think of the estimator in equation 4.25 as an estimator of $E(Y|\mathcal{X})$.

The estimator for conditional quantile would be as the usual definition of quantile.

$$Q(\mathcal{X}, q) = \inf\{y \in \mathbb{R}, \mathcal{F}_{Y|\mathcal{X}} \geq q\} \tag{4.14}$$

So, to estimate this conditional quantile, we need to estimate the conditional CDF defined in 4.13. This can be achieved using the methods of estimating empirical density functions. The estimator is based on step functions.

$$\hat{\mathcal{F}}_{Y|\mathcal{X}}(y, \mathcal{X}) = \frac{\sum K(h^{-1} d_{u,\gamma}(\mathcal{X}, \mathcal{X}_i)) \delta_g(y, y_i)}{\sum K(h^{-1} d_{u,\gamma}(\mathcal{X}, \mathcal{X}_i))} \tag{4.15}$$

This estimate is simply average of identity function, where, the function $\delta_g(y, y_i)$ controls the smoothness of the estimate. This can be defined with a different bandwidth controlling parameter g.

$$\delta_g(y, y_i) = \left\{ \begin{array}{ll} 0 & \text{for } y \leq y_i - g \\ 1 & \text{for } y > y_i + g \end{array} \right\} \tag{4.16}$$

We can use the estimator of conditional cdf to get the estimate of the the quantile

$$\hat{Q}(\mathcal{X}, q) = \inf\{y \in \mathbb{R}, \hat{\mathcal{F}}_{Y|\mathcal{X}}(y) \geq q\} \tag{4.17}$$

We now study the convergence properties of the conditional quantile estimator. First, we need show that the estimator is a consistent one.

Lemma 9. *Under the assumption of continuous conditional density exists for* $(Y|\mathcal{X})$*, and under the conditions of lemma 14,* $\forall q \in (0, 1)$*,*

$$\lim_{n\to\infty} \hat{Q}(\mathcal{X}, q) = Q(\mathcal{X}, q), \ a.co \tag{4.18}$$

Proof. This lemma uses the point wise convergence result of the conditional cdf.

Lemma 10. *Under the conditions of lemma 9, for a fixed real* y*, the estimator of the conditional cdf converges point wise to the true cdf.*

$$\lim_{n\to\infty} \hat{\mathcal{F}}_{Y|\mathcal{X}}(y) = \mathcal{F}_{Y|\mathcal{X}}(y) \tag{4.19}$$

For a proof of this lemma, we refer to the text by Ferraty [20].

Now, the continuity conditions of the lemma ensures that the conditional cdf $\hat{\mathcal{F}}_{Y|\mathcal{X}}(y)$ is continuous and strictly increasing. So, the inverse function exists and also continuous. Continuity defined at a point $Q(\mathcal{X}, q)$ can be written as,

$$\forall q \in (0, 1), \forall \epsilon > 0, \exists \delta(\epsilon) > 0, |\hat{\mathcal{F}}_{Y|\mathcal{X}}(y) - \hat{\mathcal{F}}_{Y|\mathcal{X}}(Q(\mathcal{X}, q))| \leq \delta(\epsilon) \implies |y - Q(\mathcal{X}, q)| \leq \epsilon \tag{4.20}$$

Replacing y with $\hat{Q}(\mathcal{X}, q)$ and using probability,

$$\forall q \in (0,1), \forall \epsilon > 0, \exists \delta(\epsilon) > 0,$$
$$|\hat{\mathcal{F}}_{Y|\mathcal{X}}(\hat{Q}(\mathcal{X},q)) - \hat{\mathcal{F}}_{Y|\mathcal{X}}(Q(\mathcal{X},q))| \leq \delta(\epsilon)$$
$$\implies |\hat{Q}(\mathcal{X},q) - Q(\mathcal{X},q)| \leq \epsilon$$

So, the events can be written inside a probability and using the lemma 10,

$$\forall q \in (0,1), \forall \epsilon > 0, \exists \delta(\epsilon) > 0, P(|\hat{Q}(\mathcal{X},q) - Q(\mathcal{X},q)| \leq \epsilon) \tag{4.21}$$
$$\leq P(|\hat{\mathcal{F}}_{Y|\mathcal{X}}(\hat{Q}(\mathcal{X},q)) - \hat{\mathcal{F}}_{Y|\mathcal{X}}(Q(\mathcal{X},q))| \leq \delta(\epsilon)) \tag{4.22}$$
$$= |\mathcal{F}_{Y|\mathcal{X}}(Q(\mathcal{X},q)) - \hat{\mathcal{F}}_{Y|\mathcal{X}}(Q(\mathcal{X},q))| \leq \delta(\epsilon) \tag{4.23}$$

The result follows as $\mathcal{F}_{Y|\mathcal{X}}(Q(\mathcal{X},q)) = \hat{\mathcal{F}}_{Y|\mathcal{X}}(\hat{Q}(\mathcal{X},q)) = q$.

□

4.7 Properties of Regression Estimator

We will be studying the properties of the regression estimators proposed in detail. To review the terminology, we are considering $\mathcal{X} \in \mathcal{F}$ to be a functional random variable. The observed version of $\mathcal{X}$ over time duration T is given by $X \in \mathbb{R}^T$. While defining the envelope distance, we are estimating the envelope subspace Γ_u for a envelope dimension u such that X is normally distributed with mean $\Gamma_u \eta$. Here $\Gamma_u \in \mathbb{R}^{T \times u}$ is a semi orthogonal basis of the envelope subspace. The envelope semi metric is calculated as $d_{\Gamma,u}(\mathcal{X},\mathcal{X}') = ||\Gamma_u(X - X')||^2 = ||(X - X')||^2_W$. The

weighted norm is defined as $||X||_w^2 = X^T W X$, where $W \in R^{T \times T}$ is a weighting matrix with weights determined by $W = \Gamma_u \Gamma_u^T$.

We use a kernel function $K : \mathbb{R} \mapsto \mathbb{R}^+$ to estimate the regression estimate. Specifically, we use Gaussian kernel given by $K(z) = \frac{1}{\sqrt{2\pi}} \exp\{-z^2/2\}$. We will follow the Kernel classifications as in Ferraty [20]. Specifically, we are interested in Kernel of type II, defined by

Definition 11 (Kernel of type II)**.** *A kernel function $K(z) : \mathbb{R} \mapsto \mathbb{R}^+$ such that $\int K = 1$ is called a kernel of type II if it is supported on $[0, 1]$, the derivative K' exists on $[0, 1]$ and is bounded by negative constants.*

We will show that Gaussian Kernel is of type II for future use.

Lemma 12. *Gaussian Kernel is of type II.*

Proof. $0 < K'(z) = \frac{-z}{\sqrt{2\pi}} \exp(\frac{-z^2}{2}) \leq -0.06$ □

Another quantity we need for the properties of the estimator is small ball probability corresponding to the envelope distance. The ball around a point is given by

$$B_u(\mathcal{X}, \epsilon) = \{\mathcal{X}' \in \mathcal{F}, d_{u,\gamma}(\mathcal{X}, \mathcal{X}') \leq \epsilon\} \tag{4.24}$$

The small ball probability around $\mathcal{X}$ is represented by $\phi_{\mathcal{X}}(\epsilon) = P(\mathcal{X} \in B_u(\mathcal{X}, \epsilon))$. We borrow the concept of almost complete convergence from Ferraty [20] and mention here for completeness.

Definition 13 (Almost complete convergence)**.** *A sequence of random variables* $\{X_n\}_{n\in\mathbb{N}}$ *is said to converge almost completely to a random variable* X *if and only if*

$$\forall \epsilon > 0, \sum_{n\in\mathbb{N}} P(|X_n - X| > \epsilon) < \infty$$

We will use $\lim_{n\to\infty} X_n = X, \text{a.co.}$ to denote almost complete convergence. Details about the almost complete convergence and its properties can be found in Ferraty [20].

Given above definitions and properties, the estimator for the case of regression in 4.11 is given by

$$\hat{\mu}_{k,h,u}(\mathcal{X}) = \frac{\sum_{i=1}^{n} Y_i K(h^{-1} d_{u,\Gamma}(\mathcal{X}, \mathcal{X}')}{\sum_{i=1}^{n} K(h^{-1} d_{u,\Gamma}(\mathcal{X}, \mathcal{X}'))} \tag{4.25}$$

In the following lemma, we address the complete convergence property of the estimator $\hat{\mu}_{k,h,u}$

Lemma 14. *Under the continuity condition that the true mean function* $\mu \in C^0_{\mathcal{F}}$*, where*

$$C^0_{\mathcal{F}} = \left\{ \mu : \mathcal{F} \to \mathbb{R}, \lim_{d_{\Gamma,u}(\mathcal{X},\mathcal{X}')\to 0} \mu(\mathcal{X}') = \mu(\mathcal{X}) \right\}$$

and, with a positive ball probability for $\mathcal{X}$*, given by*

$$\forall \epsilon > 0, P(\mathcal{X} \in B_u(\mathcal{X}, \epsilon)) = \phi_{\mathcal{X}}(\epsilon) > 0$$

and, the kernel K *is of type II given by 11 and the bandwidth parameter* h *is a positive sequence satisfying*

$$\lim_{n\to\infty} h = 0$$

and

$$\lim_{n\to\infty} \frac{\log n}{n\phi_{\mathcal{X}}(h)} = 0$$

and, the response variable Y satisfies

$$\forall m \geq 2, E(|Y^m||\mathcal{X}) < \sigma_m(\mathcal{X})$$

with a $\sigma_m(\mathcal{X})$ continuous at $\mathcal{X}$, then,

$$\lim_{n\to\infty} \hat{\mu}(\mathcal{X}) = \mu(\mathcal{X}), \textit{(almost completely)}$$

Proof. We first need to establish some bounds for the denominator in our estimator. We revisit the following lemma from Ferraty [20].

Lemma 15. *If K is a kernel of type II and if $\phi_{\mathcal{X}}(.)$ satisfies*

$$\exists M > 0, \exists \epsilon_0, \forall \epsilon < \epsilon_0, \int_0^{\epsilon} \phi_{\mathcal{X}}(v)dv > M\epsilon\phi_{\mathcal{X}}(\epsilon) \tag{4.26}$$

then, there exists non negative reals M_1 and M_2, such that

$$M_1\phi_{\mathcal{X}}(h) \leq EK(\frac{d_{u,\Gamma}(\mathcal{X},\mathcal{X}')}{h}) \leq M_2\phi_{\mathcal{X}}(h) \tag{4.27}$$

Proof. We know that $K'(.)$ exists, so we can write $K(t) = K(0) + \int_0^t K'(v)dv$. We denote $\nu_{u,h}$ as the measure induced by $h^{-1}K(d_{u,\Gamma}(.,.))$. So, we write

$$EK(\frac{d_{u,\Gamma}(\mathcal{X},\mathcal{X}')}{h}) = \int_0^1 K(t)d\nu_{u,h}$$

Using the expansion mentioned above, we can write

$$\begin{aligned} EK(\frac{d_{u,\Gamma}(\mathcal{X},\mathcal{X}')}{h}) &= \int_0^1 K(0)d\nu_{u,h} + \int_0^1 (\int_0^t K'(v)dv)d\nu \\ &= K(0)\phi_{\mathcal{X}}(h) + \int_0^1 (\int_0^t K'(v)I_{[v,1]}(t)dv)d\nu_{u,h} \\ &= \int_0^1 K'(v)P(v \leq \frac{d_{u,\Gamma}(\mathcal{X},\mathcal{X}')}{h} \leq 1)dv \\ &= -\int_0^1 K'(v)\phi_{\mathcal{X}}(hv)dv \end{aligned}$$

Using 4.26, for $h < \epsilon_0$, for a constant M_1

$$EK(\frac{d_{u,\Gamma}(\mathcal{X},\mathcal{X}')}{h}) \geq M_1\phi_{\mathcal{X}}(h)$$

For upper bound, we use the small ball probability as

$$P(\mathcal{X} \in B_u(\mathcal{X},h) = E(I_{[0,1]}(\frac{d_{u,\Gamma}(\mathcal{X},\mathcal{X}')}{h}))$$

As $K(.)$ is bounded in support $[0,1]$, so with $M_2 = \sup_{t\in[0,1]} K(t)$, we can bound as

$$EK(\frac{d_{u,\Gamma}(\mathcal{X},\mathcal{X}')}{h}) \leq M_2\phi_{\mathcal{X}}(h)$$

□

Now, for $i = 1, \cdots, n$, we define

$$\Delta_i = \frac{K(h^{-1}d_{u,\Gamma}(\mathcal{X},\mathcal{X}_\rangle))}{EK(h^{-1}d_{u,\Gamma}(\mathcal{X},\mathcal{X}_\infty))}$$

From the lemma above and with the regularity conditions assumed, the denominator is positive. Following similar development as in Ferraty [20] define following

two quantities

$$\hat{\mu}_1(\mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i \tag{4.28}$$

$$\hat{\mu}_2(\mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n}Y_i\Delta_i \tag{4.29}$$

We borrow two lemmas from Ferraty [20], one regarding the properties of $\hat{mu}_1(\mathcal{X})$ and another about the properties of almost complete convergence.

Lemma 16. *A sequence* $\{u_n\}$ *with* $\lim_{n\to\infty} u_n = 0$, $X_n = O_{a.co.}(u_n)$, *and* $\lim_{n\to\infty} = l_Y$, *where* l_Y *is a real constant, then*

(i) $X_nY_n = O_{a.co.}(u_n)$

(ii) $\frac{X_n}{Y_n} = O_{a.co.}(u_n)$

And the following lemma is about the convergence properties of $\hat{\mu}_1(\mathcal{X})$.

Lemma 17. *Under the conditions of 14,*

$$\hat{\mu}_1(\mathcal{X}) - 1 = O_{a.co.}(\sqrt{\frac{\log n}{n\phi_{\mathcal{X}}(h)}}) \tag{4.30}$$

We refer to Ferraty [20] for proofs of these two lemmas. We use following decomposition,

$$\begin{aligned}\hat{\mu}(\mathcal{X}) - \mu(\mathcal{X}) = &\frac{1}{\hat{\mu}_1(\mathcal{X})}\{(\hat{\mu}_2(\mathcal{X}) - E\mathcal{X}))\} \\ &-\frac{\mu(\mathcal{X})}{\hat{\mu}_1(\mathcal{X})}\{\hat{\mu}_1(\mathcal{X}) - 1\}\end{aligned}$$

A combination of this decomposition with above two lemmas completes the proof.

□

4.8 Computing Envelope Distance

The envelope distance for functional data mentioned in (4.6) requires computation of the semi orthogonal basis vectors for the envelope subspace. This is equivalent to estimating the basic matrix Γ from (4.4). As we have a parametric distribution assumed on the data, typical approach here is to use maximum likelihood estimators for Γ. We present the partially maximized log likelihood from Cook [14]

Lemma 18. *If the observed data follows the generative process of (4.4), then for a fixed u, the maximum likelihood estimator is given by*

$$argmax_{\Gamma u}(\Gamma) = -(n/2)\log|\Gamma^T S_x \Gamma| - (n/2)\log|\Gamma^T T_x^{-1}\Gamma| - (n/2)\log|T_x| - nr/2 \tag{4.31}$$

Where, S_x denote the sample covariance and T_x denote the raw second moment.

Proof. This follows from writing down the likelihood and substituting the values of S_x and T_x and fixing $\eta = \Gamma^T \bar{X}$, where $\bar{X}$ denote the sample mean for X. For details, we refer to Cook [14] □

Even though we have a close form of the objective function, however it is not convex in Γ. An efficient algorithm called 1D algorithm was proposed by Cook and Zhang Cook and Zhang [15]. However for applicability of their algorithm, they minimize a generic objective function of the form

$$J(\Gamma) = \log|\Gamma^T M \Gamma| + \log|\Gamma^T (M+U)^{-1}\Gamma| \tag{4.32}$$

Where, M and U are any symmetric and positive definite matrices. Our formulation is a special case of this objective function. The following lemma connects the two.

Lemma 19. *If we use* $M \equiv S_x = (X - \bar{X})(X - \bar{X})^T$ *and* $U \equiv \bar{X}\bar{X}^T$*, then the minimizer* $\hat{\Gamma}$ *of the objective function (4.32) spans the mean envelope for model (4.4)*

Proof. We notice in (4.31), $T_x = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$. By simple algebra, the sample covariance can be decomposed as $S_x = T_x - \bar{X}\bar{X}^T$. Now substituting $M = S_x$ and $U = \bar{X}\bar{X}^T$, we see the objective function (4.32) becomes equivalent to that of (4.31). Hence the minimizer of $J(\Gamma)$ would be the maximum likelihood estimator of the model (4.4). □

In their paper Cook and Zhang [15], Cook and Zhang proposed 1D algorithm for solving the objective function $J(\Gamma)$. They have also shown $\sqrt{n}$ consistency of the estimator achieved by the algorithm. They have also demonstrated that using a combination of 1D algorithm and the Grassmanian Manifold optimization library sg_min, they achieved more accurate result. For our purpose, we use the library developed by Zhang Zhang [53] in MATLAB for estimating the envelope basis.

The estimated envelope distance is used to calculate the projection in to the envelope subspace and plugged into the kernel. The generalize the use of envelope distance in a functional learning problem in the algorithm envelopeFDA.

4.9 Comparison with relevant methods

There is a large amount of literature on learning from functional data. However, as we are using the non parametric approach using the semi metric definitions, we limit our comparisons within methods using other semi metrics. The most common semi metrics in this area are principal component and metrics based on basis functions like Fourier or spline and their derivatives.

4.9.1 Simulation Study

We compare the performance of the envelope distance based learning algorithms with other common semi metrics. We perform our comparisons in both classification and regression setting.

For classification, we generate two classes with different envelope structure. The semi ortho-normal basis for each class was generated using orthogogonalizing a matrix with elements taken from a normal distribution. The η for both bases was kept as one vector. So the only difference between two classes arise from different basis functions. We perform non parametric classification using several distances. Other distance were L_p, PC, Fourier, Derivative distance and we also compare with KNN smoothing with euclidean distance.In figure 4.2 we show boxplots of classification accuracy after 50 repetitions of the experiment. The accuracy is calculated using a 80-20 random split of the simulated data.

For regression, we generate the predictors with envelope structure, using the similar generation method as above. The response was generated from the map

$||X^T\Gamma||_2^2$, so that the response is related to the envelope structure. It is to be noted that any other map $g : \mathbb{X} \mapsto \mathbb{R}$ would also be interesting to study as possible data generation mechanism. We compare the mean squared errors after fitting non parametric regression models using a similar 80-20 split of the data. The comparisons were done 50 times and we report the boxplot of the mean squared errors in figure 4.2

In our real data analysis we show the comparison of envelope distance with PC distance. We use the data set *tecator* reporting protein and fat content of pieces of meat. This data set is described in chapter 1 1.4.1. We plot the two components of the envelope and PC representations and color them with the content of fat and protein in figure 4.1. It can be seen that the envelope components seem to split the data by their protein and fat content while this split is not clear in PC distance.

4.9.2 Bandwidth selection

The purpose of this simulation experiment is to study the effect of data generation assumptions on the performance of the envelope regression. For this experiment we only focus on functional regression with envelope distance. There are several important questions here. As our method is non parametric in nature, the selection of bandwidth parameter plays a crucial role here.

The optimal bandwidth we are choosing is given by:

$$h^* = arg\min_h \hat{MSPE}(\hat{Y}|X)_h \tag{4.33}$$

Where, we calculate estimate of mean squared prediction error using a split of the data and fitting the non parametric regression for a fixed bandwidth. We then predict on the holdout set and repeat this process for multiple times to get an estimate of the MSPE corresponding to a specific bandwidth. We then repeat the process for a grid of bandwidth parameters and choose the parameter with lowest estimated MSPE.

We use the following data generation methods for this study.

We assume the data has an envelope subspace of dimension $d = 2$. We generate the components of the envelope basis in the same way as 4.9.1. Given the basis of the envelope subspace as (z_1, z_2), we generate the response by using different link functions described as follows.

1. Linear: $\mu(z_1, z_2) = \beta_0 + \beta_1 z_1 + \beta_2 z_2$

2. Non linear additive: $\mu(z_1, z_2) = exp(-2 * z_1 * z_2) + 6 * \tan(7 * z_1/2\pi) - 5 * \sin((30 * z_2)/2\pi)$

3. Non Linear non additive: $\mu(z_1, z_2) = exp(2 * z_1 * z_2) + 6 * \tan(7 * z_1/2\pi) - 5 * \sin((30 * z_2)/2\pi)$

We then generate the response using the model $Y(t) = \mu(t) + E(t)$, where $E(t)$ is independent standard normal noise at time point t. We show the predictors and responses in figure 4.3 and 4.4

In figure 4.5 we show side by side boxplots of mean squared errors using different semi metrics in functional non parametric regression.

4.10 Analysis of Arctic Oscillation

Changes in the Arctic Climate over the past decades have been studied and linked with many important weather phenomenon of recent times. The phase alignment of inter annual sea level pressure in arctic and southern regions and its relationship with drift in the sea ice has been studied in Power and Mysak [41], Walsh et al. [49]. Significant changes in plant ecology, physiological processes and their effect in human community is explored in Chapin III et al. [11] and Hinzman et al. [25].

The use of empirical Eigenvectors to study sea level pressure has been studied for long Kutzbach [31]. The idea has been formalized in Thompson and Wallace [48] to define the Arctic Oscillation index as the leading empirical orthogonal function of Arctic sea level pressure data. This is correlated with the surface air temperature fluctuations. Conceptually this is similar to the North Atlantic Oscillation index, however, geographically it is defined for the arctic region. As NAO has been used to study relationships between sea level pressure and other changes of the climate Gillett et al. [22]. We will be following a similar study on relationship of AO with Sea Ice Movement and Sea Ice concentration Rigor et al. [46]. Similar works studying relationship of AO with global warming is explored in Fyfe et al. [21], Chambers and Ogle [10]. Studies on sea ice cover Comiso et al. [13] and explores its accelerated decline. The recent decrease in AO has been studied in Overland and Wang [39]. Use of AO to predict and explain effects of climate change has been studied in Aanes et al. [2].

Sea level pressure is an important quantity for climate scientists, which is usually used to describe and explain many weather phenomenon. Sea level pressure

data is observed over a grid at frequent time intervals. We have collected the SLP series for northern hemisphere starting from 20N to 90N from the database IRIDL. The data is collected on a spatial grid of 2.5 x 2.5 degrees. We have taken monthly summaries of the anomaly of the series, by centering it with average of the yearly observations. The data is collected from January 1979 to January 2001.

It is common practice in atmospheric sciences to summarize this grid data of SLP into a single time series for the ease of explanation. This is generally done with the help of Empirical Orthogonal Function. It is generally accepted that the leading EOF of the SLP in a pre specified grid captures the most fraction of the explained variance. The grid observations are then projected onto this leading EOF to construct the Arctic Oscillation series. For details of constructing AO, we refer to CPC.

4.10.1 Reconstruction of Arctic Oscillation Series

In the methodology section of the Arctic Oscillation documentation (`http://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/history/method.shtml`) by the Climate Prediction Center it is explained that Empirical Orthogonal Function (EOF) was applied to the covariance matrix of the monthly mean 1000 hPa height anomalies polewards 20 degree North (or South) through the period of 1979 to 2000, collected on a grid of $(2.5^o \times 2.5^o)$. It is also mentioned that to ensure equal area weighting for the covariance matrix, the gridded data was weighted by the square root of the cosine of the latitude.

The use of Empirical Orthogonal Function to study weather related data is

an established practice in the field of statistical weather prediction Lorenz [33]. In a weather data set if there are time series of of length T being observed over p locations, then we can tabulate the data into a matrix X of dimension $T \times p$. Here p can be a list of weather stations or a grid over an area where the data is being collected. The idea behind Empirical Orthogonal Functions is to expand each time series in terms of optimally weighted functions over location.

Without loss of generality, it is assumed that the series $X(t)$ have mean zero. This can be easily achieved by removing the mean from each series. This series is called anomaly series in climate literature. The location dependency is represented in the empirical covariance estimate of $\hat{\Sigma} = S = X^T X$. In EOF analysis, this covariance matrix is decomposed into Eigen values and vectors.

$$S\phi_i = \lambda_i \phi_i \tag{4.34}$$

Where the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$.

So, this decomposition is also known as principal component analysis in many other fields. This allows for a representation of the time series as

$$X(t) = \sum_{i=1}^{p} \alpha(t)\phi_i \tag{4.35}$$

The expansion coefficients $\alpha(t)$ are the principal components and the Eigen vectors ϕ_i are the loading for the locations. Because these are ordered by the decreasing order of eigenvalues, truncated sum of eq4.35 would increase with inclusion of more components. The components explain the variance of X in a decreasing order. These properties about principal component analysis are very well known. In the

original work Lorenz [33] proposing the EOF on weather data, a full derivation of the method is presented starting from a slight variation of the problem.

The eigen vector corresponding to the largest eigen value λ_1 is called the first leading mode from the EOF analysis. This vector indicates the loading pattern of the location time series. The projection of the data matrix into the direction of the loading pattern would product Arctic Oscillation (AO).

There are various applications of the loading patterns of AO in the climate literature. However, there are relationship between several other modes often used in the climate data analysis. The work of Monahan et al. [34] relates the EOF modes with dynamic modes, kinematic degrees of freedom, etc. So, various modes can not be interpreted individually independent of other modes.

The sample covariance estimator can be quite noisy, especially when there are limited number of time points available, which is often the case. The work by North et al. [38] studies the effect of sampling error in the noise arisen in the Eigen vectors (empirical orthogonal functions or EOFs) in detail and suggests a grid based weighting as a possible remedy. In the definition of AO, it is mentioned that the covariance matrix is spatially weighed with weights proportional to square root of cosine of latitude.

In general the weighting matrix is a square matrix W of dimension $p \times p$. The estimate of the covariance matrix from an individual time point is given by $x^T W x$. So, the estimate for the location covariance matrix becomes:

$$\frac{1}{n} \sum_{t=1}^{T} x_t^T W x_t \tag{4.36}$$

Where the weighting matrix $W = [w_{ij}]$ indicates weights for location (ij). For

simplicity, we will consider the diagonal weighting only. The diagonal elements will be square root of cosine of latitudes for corresponding locations. The estimation for the eigenvalues will be quite similar. The data is transformed as $X' = XW^{1/2}$ and the covariance matrix becomes $S' = (W^{1/2})^T S W^{1/2}$. Or, it is also simply changed in the projection into the first principal component as

$$X'_{proj} = \frac{XW\phi_1}{\phi_1^T W \phi_1} \tag{4.37}$$

Where the original Eigenvectors ϕ_i were calculated from the un-weighted data. Several other methods on implementing spatial weighting and iterative estimation of EOF is considered in Baldwin et al. [3].

For comparison, we consider both cases here with and without spatially weighting and unweighted versions of the covariance matrix. In figure 4.6, we show the EOF first component for two cases on the grid of locations.

It has been a common approach in analyzing the sea level pressure data is to consider summer and winter seasons separately. In figure 4.7 and 4.8 we are showing the contour plots of the SLP with location in long or lat for summer and winter months. It is clear that the two seasons have quite different pressure anomaly as expected.

The data on sea level pressure does have clear seasonal component which is evident from the time series plot of the data at selected locations. In figure 4.9 we are showing time series plots of SLP for 22 months at 20N and at the north pole. The yearly sections has been separated out to show the seasonality of data.

We downloaded the original Arctic Oscillation series from the NOAA website (`https://www.ncdc.noaa.gov/teleconnections/ao/`) on methodology of AO.

We compared the version of the series within 1979 to 2000 with the reconstructed series from weighted EOF analysis. In figure 4.10 we show all three series. The spatially weighted version of the reconstruction seems to match with the original AO satisfactorily, and we decide to go ahead with this version of reconstruction in future analysis.

We also look at the linear fit on the scatterplot of the original AO and the reconstructed AO in figure 4.11. The two lines shown indicates the fit with and without intercept. The results of the linear fit is shown in table 4.1. The R squared value for the linear fit with intercept was 0.918. As the estimated slope is very high and intercept is low enough, we consider the reconstructed AO as a good proxy for the original AO series.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1277	0.0177	7.22	0.0000
Slope	0.9562	0.0177	54.08	0.0000

Table 4.1: Regression summery for Original AO vs Reconstructed AO (spatially weighted)

There has been a volume of research linking Sea Level Pressure with other atmospheric events. It is well understood that SLP has explanatory power over other quantities related to Sea Ice. Several works have shown relationship between Sea Ice Concentration with SLP]. There has also been works on understanding the movements of Sea Ice Sheets and its relations with SLP. In most of these studies, the AO series is used as the predictor representing SLP.

We collected the Sea Ice Concentration data from (`http://nsidc.org/data/`

docs/noaa/g02135_seaice_index/index.html) the NSIDC. The monthly averages of SIC has been reported and they have been summarized over all the grid points of observation. We have constructed the monthly anomalies series by centering the series with the averages. The monthly SIC series is shown at 4.14. Here monthly anomalies are calculated by subtracting monthly means from the monthly series and presented as a ratio of monthly means.

The yearly anomaly SIC series is shown in figure 4.12, where the series has been standardized by dividing by standard deviation.

There has been a few missing observation of SIC. We had to remove the corresponding months from the SLP dataset as well. The number of grid points for SLP data is 4177. We have considered this dataset as a spatially variable series for each time point. We are interested in the relationship between SLP and SIC. In fig 4.13, we show the SLP series plotted against locations for each month.

To get the Empirical Orthogonal Functions, we construct the covariance matrix from the SLP series. We then use decompose it into eigen vectors and pick the one corresponding to the largest eigen value as the leading EOF. We project the SLP into this EOF, by taking a linear combination of the grid locations. Thus we have reconstructed the equivalent of AO in our dataset. Off course, we can also use the original AO series within our time interval.

We report the results of regression models with reconstructed AO and SIC in 4.15. The R squared is close to zero and the p value for the F statistic is 0.71. So this does not show sufficient evidence for linear relationship between the two quantities.

There has been little amount of work applying functional data analysis techniques on climate data. Although it was mentioned as a possible application area in J. O. Ramsay [28] and a brief case study was presented at Ramsay and Silverman [44], we found no detailed application example of functional data analysis models in climate related data. However, as the data we are considering is spatio temporal by nature, such an analysis is not inconceivable.

We used the kernel based functional regression with envelope distance with envelope dimension as 1. We show the results in 4.16. The R squared was 0.752, indicating a strong relationship between the predictor and the response.

	R2	sd(R2)	MSE	sd(MSE)	MAD	sd(MAD)	cor	sd(cor)
lmRecon	0.00	0.00	0.14	0.03	0.31	0.03	-0.07	0.12
lmOrig	0.00	0.00	0.14	0.03	0.31	0.03	-0.03	0.13
lmOrig2	0.01	0.00	0.14	0.03	0.31	0.03	-0.03	0.13
PC1	0.70	0.03	0.22	0.04	0.37	0.04	-0.00	0.13
PC2	0.59	0.05	0.19	0.04	0.36	0.04	-0.00	0.12
PC3	0.42	0.03	0.17	0.03	0.33	0.03	0.02	0.13
Envlp1	0.70	0.04	0.26	0.05	0.42	0.05	-0.07	0.20
Envlp2	0.59	0.02	0.24	0.04	0.39	0.04	0.04	0.09
Envlp3	0.49	0.04	0.25	0.05	0.40	0.04	0.02	0.12
F5	0.29	0.02	0.15	0.03	0.32	0.03	-0.04	0.11
F7	0.28	0.02	0.15	0.03	0.32	0.03	-0.03	0.10
F9	0.27	0.02	0.15	0.03	0.31	0.03	-0.02	0.10
F11	0.27	0.02	0.15	0.03	0.31	0.03	-0.02	0.10
B5	0.32	0.03	0.15	0.03	0.32	0.03	-0.03	0.11
B7	0.29	0.02	0.15	0.03	0.32	0.03	-0.02	0.11
B9	0.28	0.02	0.15	0.03	0.31	0.03	-0.02	0.10
B11	0.27	0.02	0.15	0.03	0.31	0.03	-0.02	0.10
Deriv1	0.30	0.02	0.14	0.03	0.31	0.03	-0.03	0.08
Deriv2	0.18	0.01	0.14	0.03	0.31	0.03	-0.10	0.17
Deriv3	0.33	0.02	0.15	0.03	0.32	0.03	0.02	0.08

We used the time series data of AO and the SIC, and used a split of $80 - 20$

to create a training and testing set for our models. This however does destroy the time variate structure of the data, but we can compare the methods in terms of predictive performance. We have considered linear regression, and various semi metric based kernel regression with PC, envelope, Fourier, Bspline and derivative distances of various order. We present the means and standard deviations of the metrics in table 4.10.1.

In figure 4.17, we show the time series of sea level pressure for the time period 1979 to 2000 with the predicted series using AO and Envelope. It is clear that the envelope method produces much closer approximation than using linear regression on AO.

In figure 4.18 and in figure 4.19, we show the loading patterns of AO and envelope. For AO, the loading pattern is generated using the first principal component of the matrix of location and time points of sea level pressure data. For envelope, the vector denoting the basis of the envelope subspace is considered as loading pattern. They both has been projected onto the world map within latitude up words of 20 North. Mercator projection has been used to project the rectangular grid data into spherical plot.

Bibliography

[1] *Linear Processes in Function Spaces - Theory and Applications | Denis Bosq | Springer*. URL `http://www.springer.com/us/book/9780387950525`.

[2] Ronny Aanes, Bernt-Erik Sæther, Fiona M Smith, Elisabeth J Cooper,

Philip A Wookey, and Nils Are Øritsland. The arctic oscillation predicts effects of climate change in two trophic levels in a high-arctic ecosystem. *Ecology Letters*, 5(3):445–453, 2002.

[3] Mark P. Baldwin, David B. Stephenson, and Ian T. Jolliffe. Spatial Weighting and Iterative Projection Methods for EOFs. *Journal of Climate*, 22(2):234–243, January 2009. ISSN 0894-8755, 1520-0442. doi: 10.1175/2008JCLI2147.1. URL http://journals.ametsoc.org/doi/abs/10.1175/2008JCLI2147.1.

[4] Philippe Besse and J. O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311, June 1986. ISSN 0033-3123, 1860-0980. doi: 10.1007/BF02293986. URL http://link.springer.com/article/10.1007/BF02293986.

[5] Graciela Boente and Ricardo Fraiman. Kernel-based functional principal components. *Statistics and Probability Letters*, 48(4):335–345, 2000.

[6] HervÃl' Cardot. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12(4):503–538, January 2000. ISSN 1048-5252. doi: 10.1080/10485250008832820. URL http://dx.doi.org/10.1080/10485250008832820.

[7] HervÃl' Cardot, FrÃl'dÃl'ric Ferraty, and Pascal Sarda. Functional linear

model. *Statistics & Probability Letters*, 45(1):11–22, 1999. URL http://EconPapers.repec.org/RePEc:eee:stapro:v:45:y:1999:i:1:p:11-22.

[8] HervÃl' Cardot, FrÃl'dÃl'ric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, pages 571–591, 2003. URL http://www.jstor.org/stable/24307112.

[9] P. E. Castro, W. H. Lawton, and E. A. Sylvestre. Principal Modes of Variation for Processes with Continuous Sample Curves. *Technometrics*, 28(4):329–337, 1986. ISSN 0040-1706. doi: 10.2307/1268982. URL http://www.jstor.org/stable/1268982.

[10] Frank Chambers and Michael Ogle, editors. *Climate change: critical concepts in the environment*. Routledge, London ; New York, 2002. ISBN 978-0-415-27656-6 978-0-415-27657-3 978-0-415-27658-0 978-0-415-27659-7 978-0-415-27660-3.

[11] F. Stuart Chapin III, Robert L Jefferies, James F Reynolds, Gaius R Shaver, Josef Svoboda, and Ellen W Chu. *Arctic Ecosystems in a Changing Climate: an Ecophysiological Perspective*. Elsevier Science, Oxford, 1991. ISBN 978-0-323-13842-0. URL http://www.123library.org/book_details/?id=64948. OCLC: 829460348.

[12] Dong Chen, Peter Hall, and Hans-Georg MÃijller. Single and multiple index

functional regression models with nonparametric link. *The Annals of Statistics*, 39(3):1720–1747, June 2011. ISSN 0090-5364, 2168-8966. doi: 10.1214/11-AOS882. URL http://projecteuclid.org/euclid.aos/1311600281.

[13] Josefino C. Comiso, Claire L. Parkinson, Robert Gersten, and Larry Stock. Accelerated decline in the Arctic sea ice cover. *Geophysical Research Letters*, 35(1), January 2008. ISSN 0094-8276. doi: 10.1029/2007GL031972. URL http://doi.wiley.com/10.1029/2007GL031972.

[14] R. D. Cook. Class Notes Envelope, 2014.

[15] R. Dennis Cook and Xin Zhang. Algorithms for Envelope Estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300, January 2016. ISSN 1061-8600. doi: 10.1080/10618600.2015.1029577. URL http://dx.doi.org/10.1080/10618600.2015.1029577.

[16] R. Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960, 2010. URL http://www.jstor.org/stable/24309466.

[17] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154, March 1982. ISSN 0047-259X. doi: 10.1016/0047-259X(82)90088-4. URL http://www.sciencedirect.com/science/article/pii/0047259X82900884.

[18] Winston Wei Dou, David Pollard, and Harrison H. Zhou. Estimation in functional regression for general exponential families. *The Annals of Statistics*, 40(5):2421–2451, October 2012. ISSN 0090-5364, 2168-8966. doi: 10.1214/12-AOS1027. URL `http://projecteuclid.org/euclid.aos/1359987526`.

[19] FrÃl'dÃl'ric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer series in statistics. Springer, New York, 2006. ISBN 978-0-387-30369-7. OCLC: ocm70261207.

[20] Vieu Philippe Ferraty, FrÃl'dÃl'ric. *Nonparametric Functional Data Analysis*. 2006.

[21] JC Fyfe, GJ Boer, and GM Flato. Arctic and antarctic oscillations and their projected changes under global warming. *Geophysical Research Letters*, 26 (11):1601–1604, 1999.

[22] Nathan P. Gillett, Francis W. Zwiers, Andrew J. Weaver, and Peter A. Stott. Detection of human influence on sea-level pressure. *Nature*, 422(6929):292–294, March 2003. ISSN 00280836. doi: 10.1038/nature01487. URL `http://www.nature.com/doifinder/10.1038/nature01487`.

[23] Peter Hall and Joel L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, February 2007. ISSN 0090-5364. doi: 10.1214/009053606000000957. URL `http://projecteuclid.org/euclid.aos/1181100181`.

[24] Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126, 2006. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00535.x/full.

[25] Larry D. Hinzman, Neil D. Bettez, W. Robert Bolton, F. Stuart Chapin, Mark B. Dyurgerov, Chris L. Fastie, Brad Griffith, Robert D. Hollister, Allen Hope, Henry P. Huntington, Anne M. Jensen, Gensuo J. Jia, Torre Jorgenson, Douglas L. Kane, David R. Klein, Gary Kofinas, Amanda H. Lynch, Andrea H. Lloyd, A. David McGuire, Frederick E. Nelson, Walter C. Oechel, Thomas E. Osterkamp, Charles H. Racine, Vladimir E. Romanovsky, Robert S. Stone, Douglas A. Stow, Matthew Sturm, Craig E. Tweedie, George L. Vourlitis, Marilyn D. Walker, Donald A. Walker, Patrick J. Webber, Jeffrey M. Welker, Kevin S. Winker, and Kenji Yoshikawa. Evidence and Implications of Recent Climate Change in Northern Alaska and Other Arctic Regions. *Climatic Change*, 72(3):251–298, October 2005. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-005-5352-2. URL http://link.springer.com/10.1007/s10584-005-5352-2.

[26] Ivana HorovÃą, Philippe Vieu, and JiÅŹÃ■ Zelinka. OPTIMAL CHOICE OF NONPARAMETRIC ESTIMATES OF A DENSITY AND OF ITS DERIVATIVES. *Statistics & Risk Modeling*, 20(1-4), January 2002. ISSN 2196-7040, 2193-1402. doi: 10.1524/strm.2002.

20.14.355. URL http://www.degruyter.com/view/j/strm.2002.20.issue-1-4/strm.2002.20.14.355/strm.2002.20.14.355.xml.

[27] Tailen Hsing and Randall L. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley series in probability and statistics. John Wiley and Sons, Inc, Chichester, West Sussex, 2015. ISBN 978-0-470-01691-6.

[28] C. J. Dalzell J. O. Ramsay. Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):539–572, 1991. ISSN 00359246. URL http://www.jstor.org/stable/2345586.

[29] Gareth M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432, August 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00342. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00342/abstract.

[30] I. T. Jolliffe. *Jolliffe I. Principal Component Analysis (2ed., Springer, 2002)(518s)_MVsa_.pdf*. Springer. URL http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf. bibtex: _jolliffe_????

[31] John E. Kutzbach. Empirical Eigenvectors of Sea-Level Pressure, Surface Temperature and Precipitation Complexes over North America.

Journal of Applied Meteorology, 6(5):791–802, October 1967. ISSN 0021-8952. doi: 10.1175/1520-0450(1967)006<0791:EEOSLP>2.0.CO; 2. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0450%281967%29006%3C0791%3AEEOSLP%3E2.0.CO%3B2.

[32] N. Locantore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, K. L. Cohen, Graciela Boente, Ricardo Fraiman, Babette Brumback, Christophe Croux, Jianqing Fan, Alois Kneip, John I. Marden, Daniel PeÃśa, Javier Prieto, Jim O. Ramsay, Mariano J. Valderrama, Ana M. Aguilera, N. Locantore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, and K. L. Cohen. Robust principal component analysis for functional data. *Test*, 8 (1):1–73, June 1999. ISSN 1133-0686, 1863-8260. doi: 10.1007/BF02595862. URL http://link.springer.com/article/10.1007/BF02595862.

[33] Edward N Lorenz. Empirical orthogonal functions and statistical weather prediction. 1956.

[34] Adam H. Monahan, John C. Fyfe, Maarten H. P. Ambaum, David B. Stephenson, and Gerald R. North. Empirical Orthogonal Functions: The Medium is the Message. *Journal of Climate*, 22(24):6501–6514, December 2009. ISSN 0894-8755. doi: 10.1175/2009JCLI3062.1. URL http://journals.ametsoc.org/doi/abs/10.1175/2009JCLI3062.1.

[35] Jeffrey S. Morris. Functional Regression. *Annual Review of Statistics and Its Application*, 2(1):321–359, April 2015. ISSN 2326-8298, 2326-831X. doi: 10.

1146/annurev-statistics-010814-020413. URL http://www.annualreviews.org/doi/10.1146/annurev-statistics-010814-020413.

[36] HANS-GEORG MÃIJLLER. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240, 2005. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2005.00429.x/abstract.

[37] Hans-Georg MÃijller and Ulrich StadtmÃijller. Generalized functional linear models. *The Annals of Statistics*, 33(2):774–805, April 2005. ISSN 0090-5364. doi: 10.1214/009053604000001156. URL http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1117114336/.

[38] Gerald R. North, Thomas L. Bell, Robert F. Cahalan, and Fanthune J. Moeng. Sampling Errors in the Estimation of Empirical Orthogonal Functions. *Monthly Weather Review*, 110(7):699–706, July 1982. ISSN 0027-0644. doi: 10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(1982)110%3C0699%3ASEITEO%3E2.0.CO%3B2.

[39] JE Overland and M Wang. The arctic climate paradox: The recent decrease of the arctic oscillation. *Geophysical Research Letters*, 32(6), 2005.

[40] S. Pezzulli and B. W. Silverman. Some properties of smoothed principal components analysis for functional data. *Computational Statistics*, 8:1–16, 1993.

[41] Scott B. Power and Lawrence A. Mysak. On the interannual variability of arctic seaâĂŘlevel pressure and sea ice. *Atmosphere-Ocean*, 30(4): 551–577, December 1992. ISSN 0705-5900, 1480-9214. doi: 10.1080/07055900.1992.9649455. URL http://www.tandfonline.com/doi/abs/10.1080/07055900.1992.9649455.

[42] C. Preda and G. Saporta. Clusterwise pls regression on a stochastic process. *Comput. Stat. Data Anal.*, 49(1):99–108, April 2005. ISSN 0167-9473. doi: 10.1016/j.csda.2004.05.002. URL http://dx.doi.org/10.1016/j.csda.2004.05.002.

[43] James Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Springer New York, New York, NY, 2009. ISBN 978-0-387-98184-0 978-0-387-98185-7. URL http://link.springer.com/10.1007/978-0-387-98185-7.

[44] James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Citeseer, 2002.

[45] John A. Rice. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, pages 631–647, 2004. URL http://www.jstor.org/stable/24307409.

[46] Ignatius G. Rigor, John M. Wallace, and Roger L. Colony. Response of sea ice to the Arctic Oscillation. *Journal of Climate*, 15(18):

2648–2663, 2002. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0442(2002)015%3C2648%3AROSITT%3E2.0.CO%3B2.

[47] Z. Shi. Small ball probabilities for a Wiener process under weighted sup-norms, with an application to the supremum of bessel local times. *Journal of Theoretical Probability*, 9(4):915–929, October 1996. ISSN 0894-9840, 1572-9230. doi: 10.1007/BF02214257. URL http://link.springer.com/article/10.1007/BF02214257.

[48] David W. J. Thompson and John M. Wallace. The Arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical Research Letters*, 25(9):1297–1300, May 1998. ISSN 00948276. doi: 10.1029/98GL00950. URL http://doi.wiley.com/10.1029/98GL00950.

[49] John E. Walsh, William L. Chapman, and Timothy L. Shy. Recent Decrease of Sea Level Pressure in the Central Arctic. *Journal of Climate*, 9(2):480–486, February 1996. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(1996)009<0480:RDOSLP>2.0.CO;2. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0442%281996%29009%3C0480%3ARDOSLP%3E2.0.CO%3B2.

[50] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg MÃijller. Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, June 2016. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-041715-033624. URL http://www.annualreviews.org/doi/10.1146/annurev-statistics-041715-033624.

[51] Suojin Wang, Lianfen Qian, and Raymond J. Carroll. Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika*, 97(1):79–93, 2010. URL `http://biomet.oxfordjournals.org/content/97/1/79.short`.

[52] Hulin Wu and Jin-Ting Zhang. *Nonparametric regression methods for longitudinal data analysis: [mixed-effects modeling approaches]*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J, 2006. ISBN 978-0-471-48350-2. OCLC: ocm62525265.

[53] Xing Zhang. EnvelopeAlgorithms. URL `http://stat.fsu.edu/~henry/EnvelopeAlgorithms.pdf`.

[54] Xin Zhao, James Stephen Marron, and Martin T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, pages 789–808, 2004. URL `http://www.jstor.org/stable/24307416`.

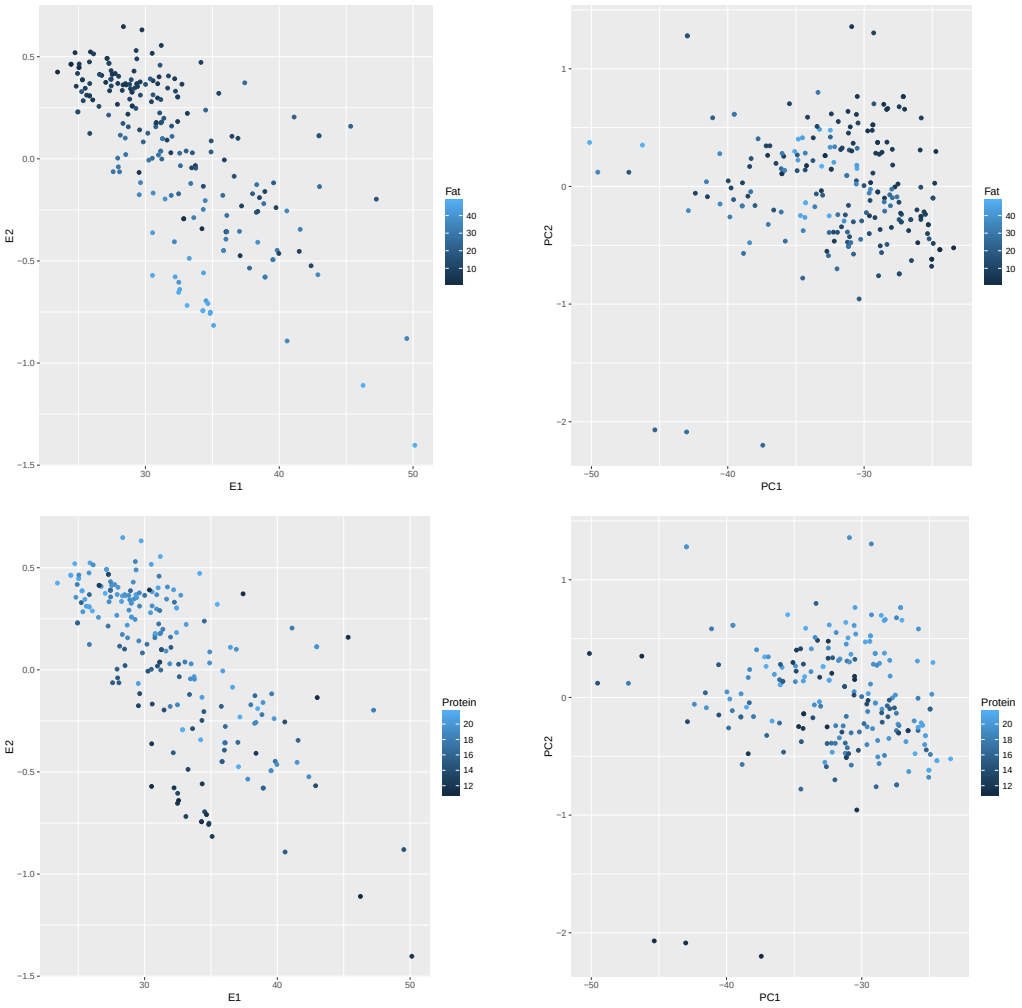


Figure 4.1: Tecator data with envelope and PC comparison.

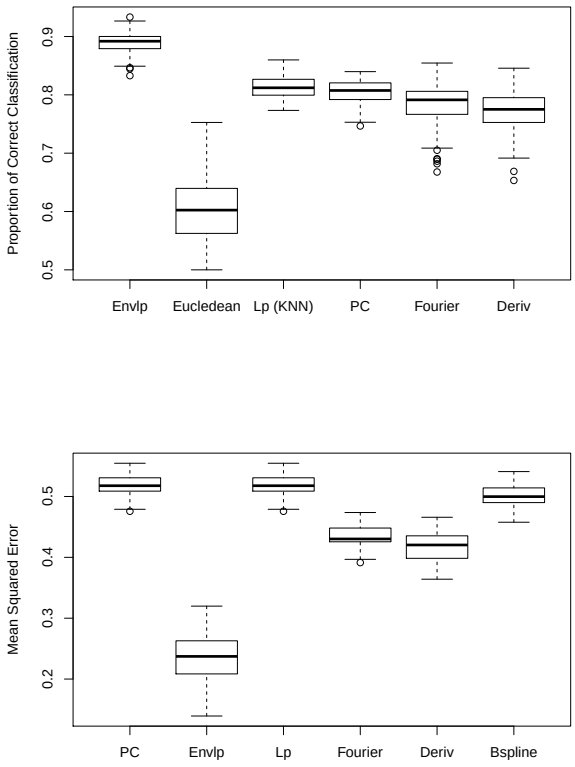


Figure 4.2: Simulation results for classification and regression

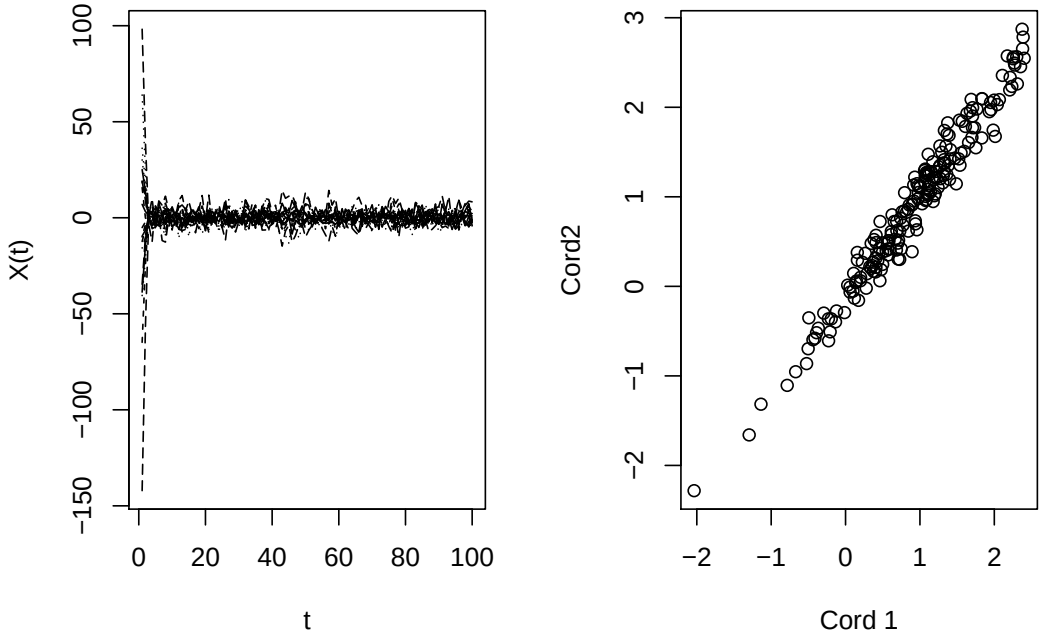


Figure 4.3: Functional predictors and projections into envelope subspace for simulation experiment.

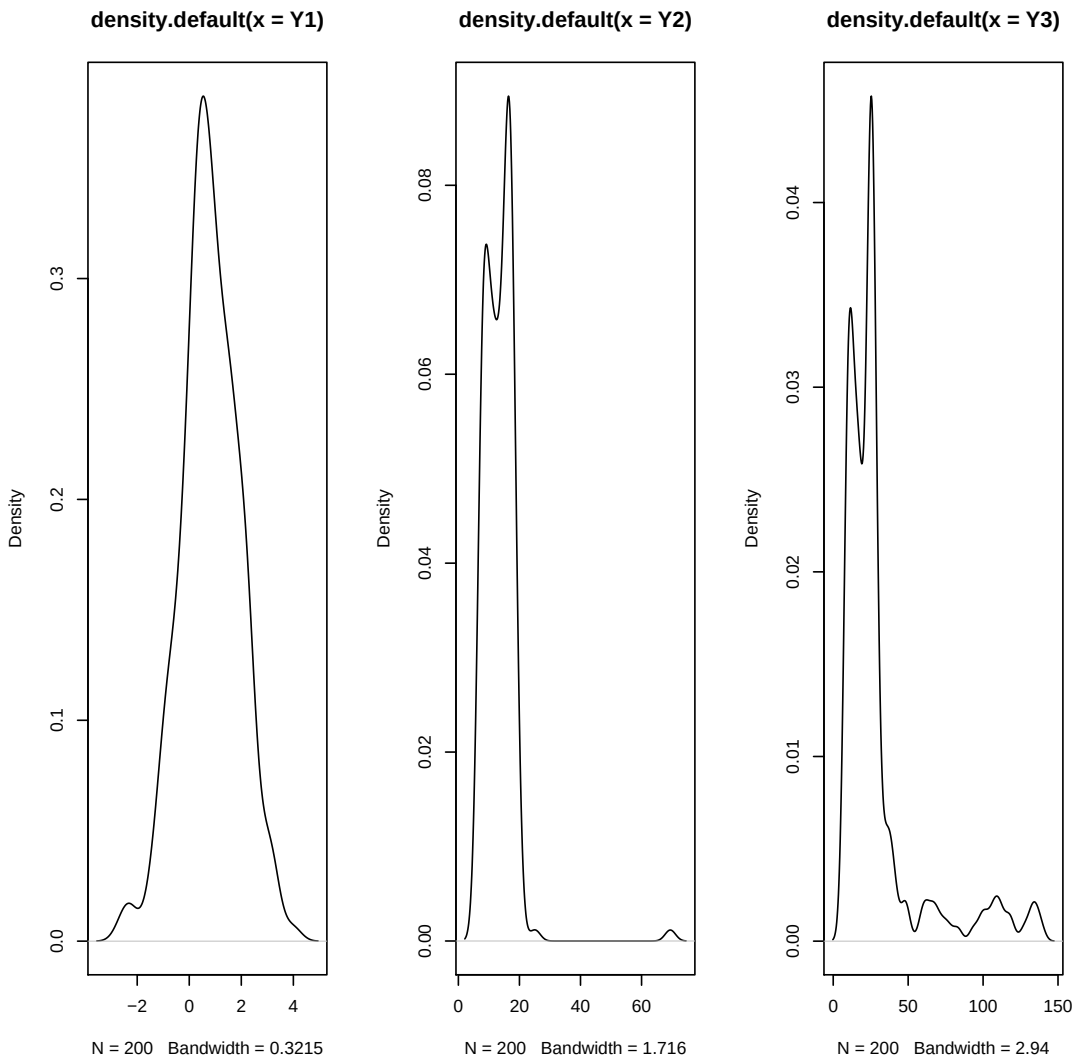


Figure 4.4: Density plots of responses using three different link functions.

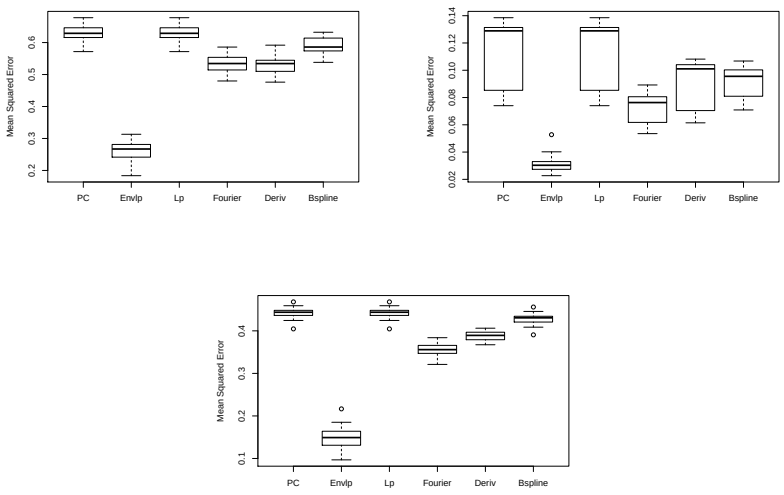


Figure 4.5: Simulation results for three regression scenarios

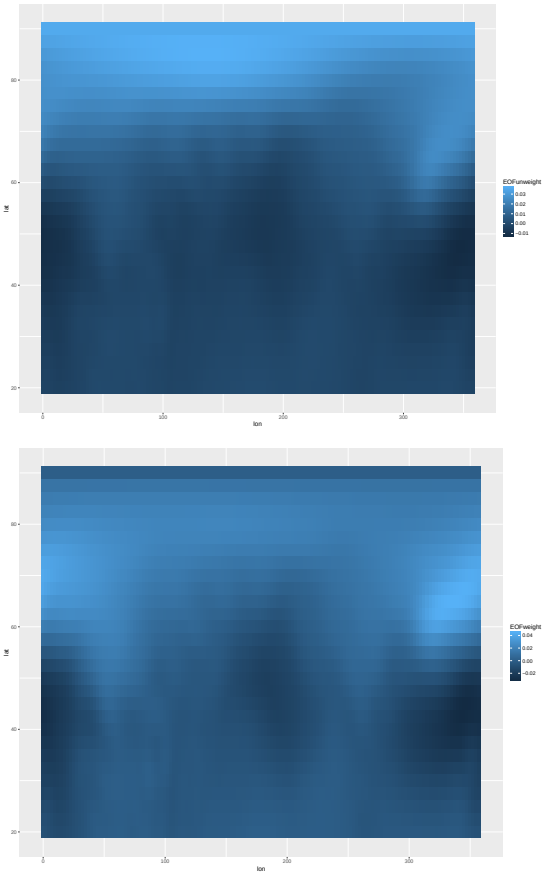


Figure 4.6: First EOF of SLP with location grid for unweighted and spatially weighted covariance.

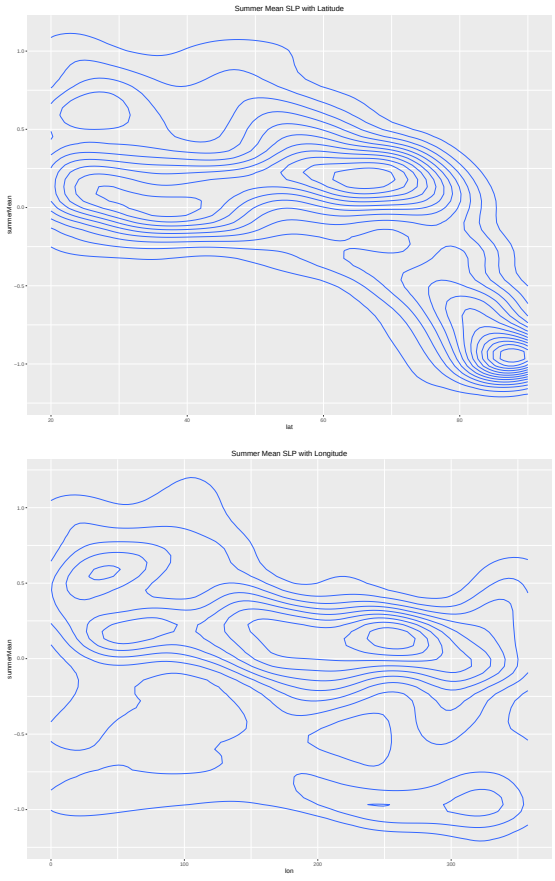


Figure 4.7: Contour plots of SLP with latitude and longitude in summer months.

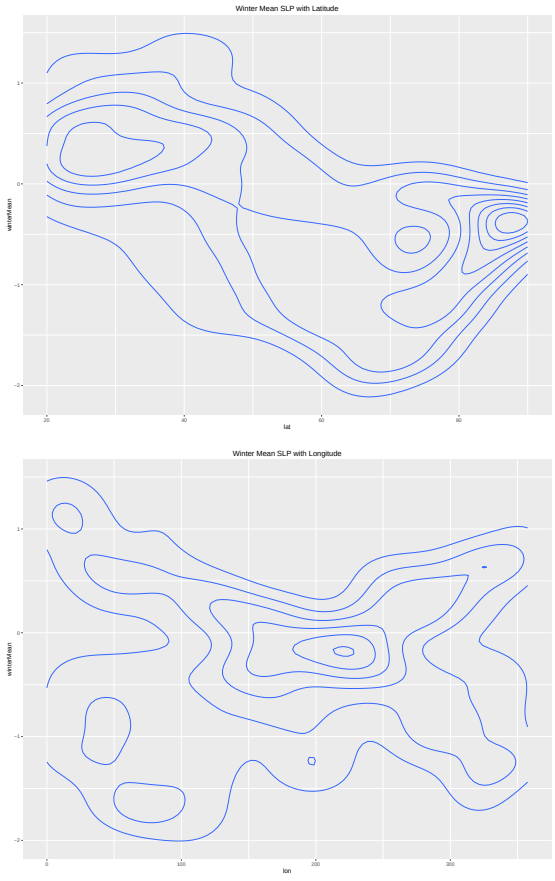


Figure 4.8: Contour plots of SLP with Latitude and longitude in winter months.

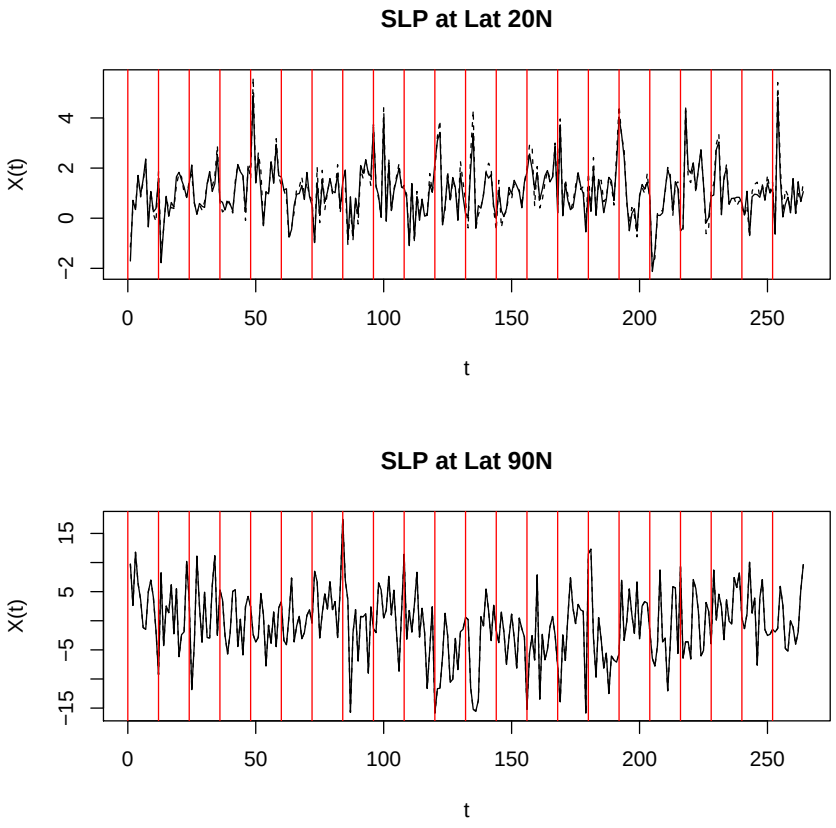


Figure 4.9: Seasonality of SLP

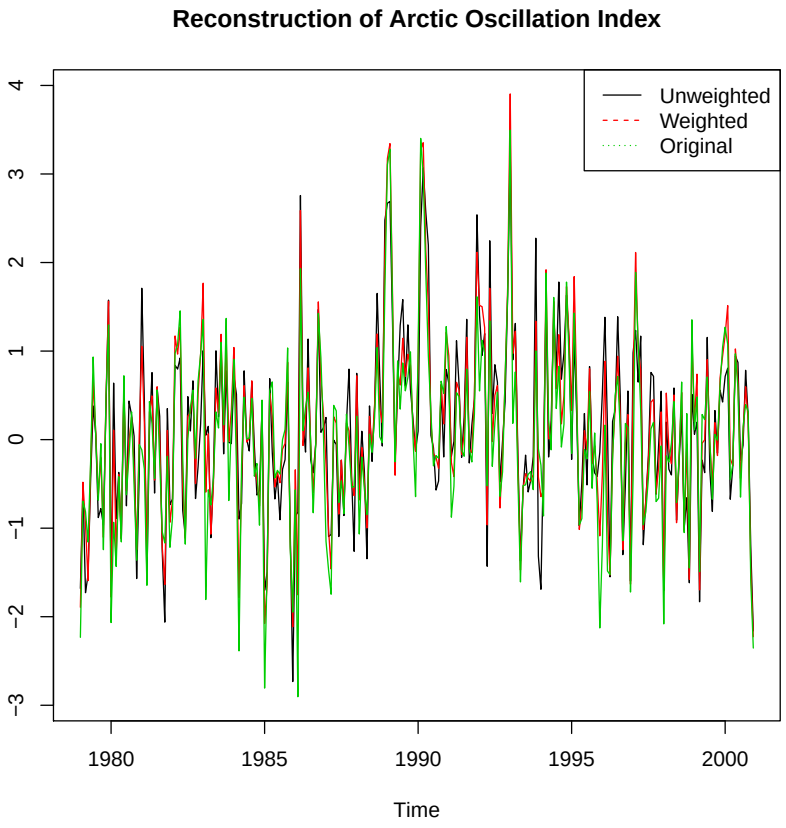


Figure 4.10: Comparison of Reconstructed AO and original AO series.

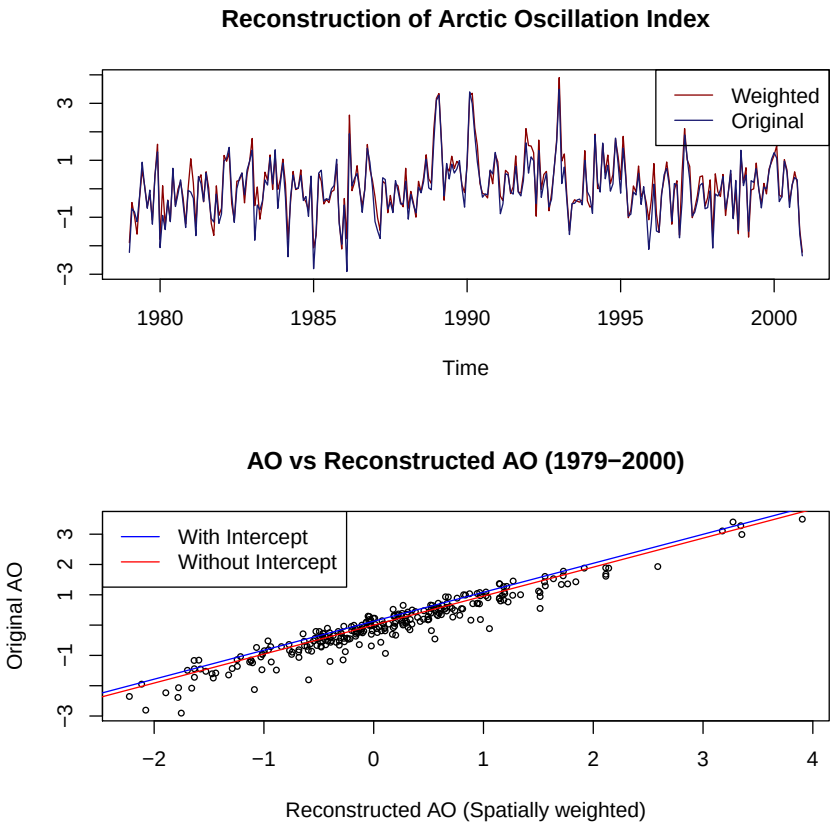


Figure 4.11: Original Vs Reconstructed AO

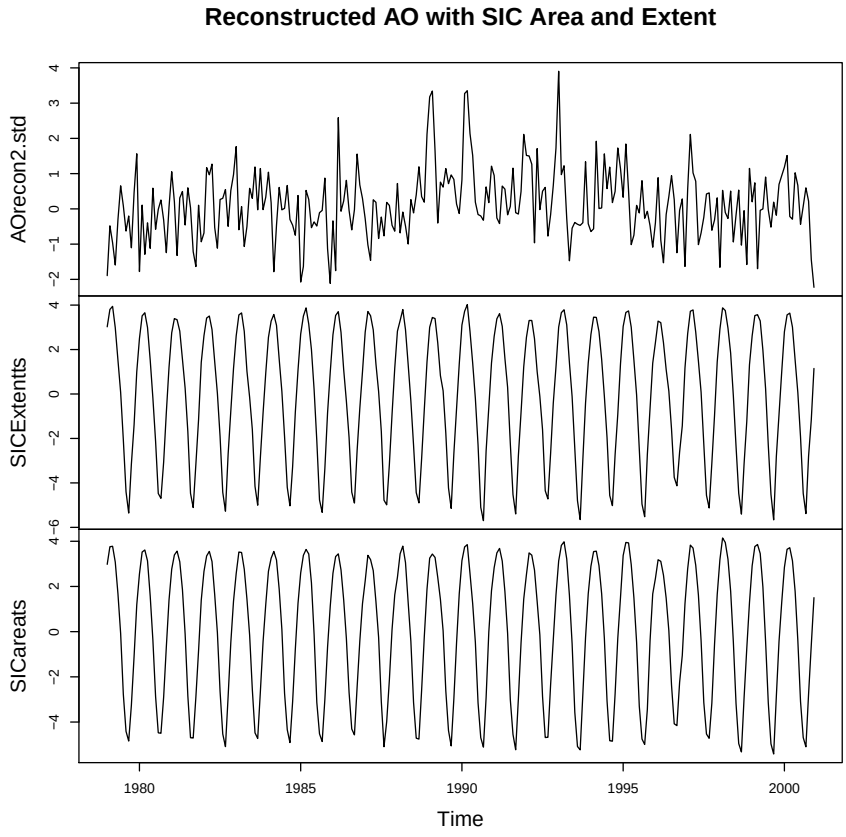


Figure 4.12: Standardized SIC yearly anomaly for sea ice area and extent.

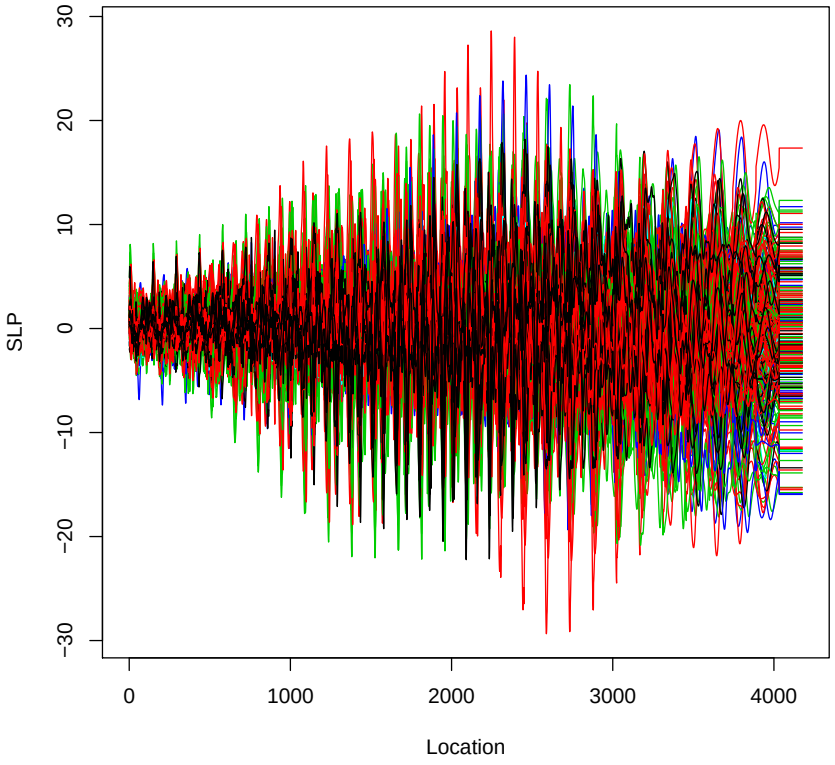


Figure 4.13: Monthly SIC at grid locations.

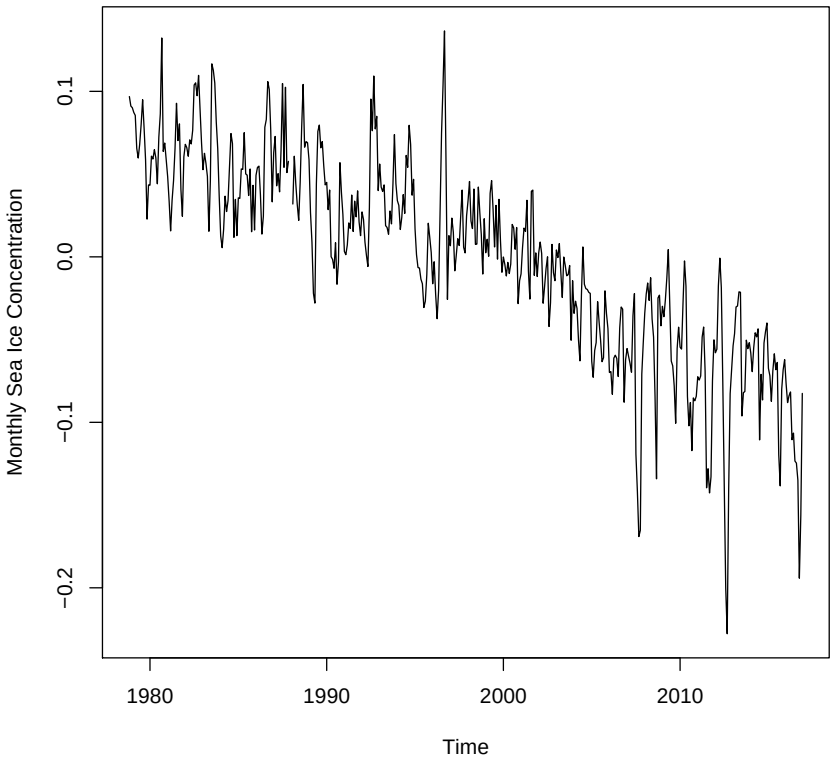


Figure 4.14: Monthly Sea Ice Concentration Anomalies

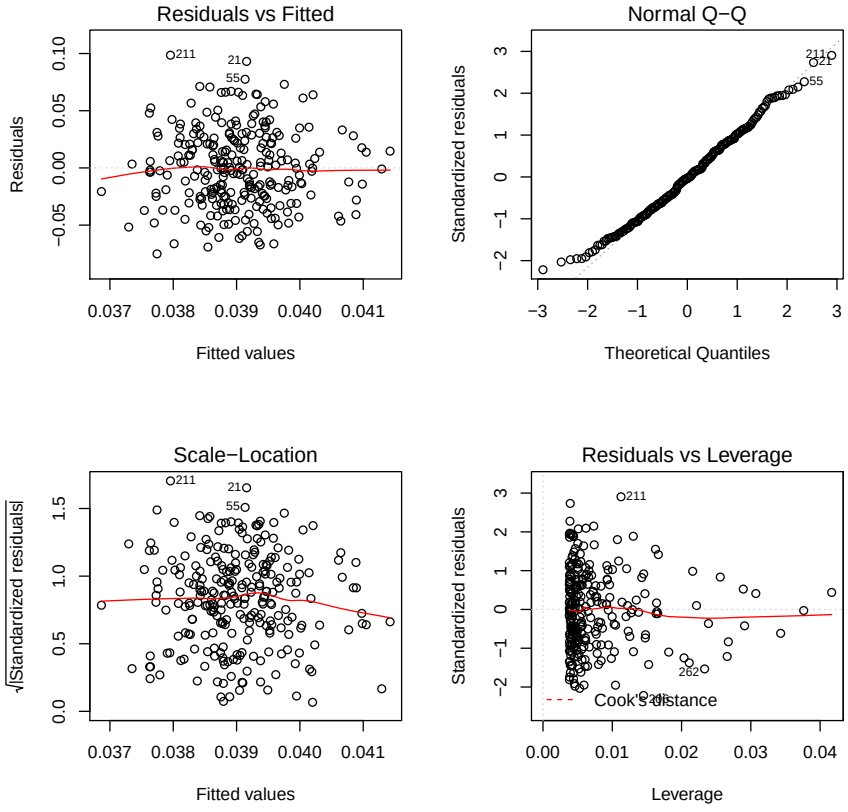


Figure 4.15: Residual and Diagnostic plots for regression of SIC on reconstructed AO

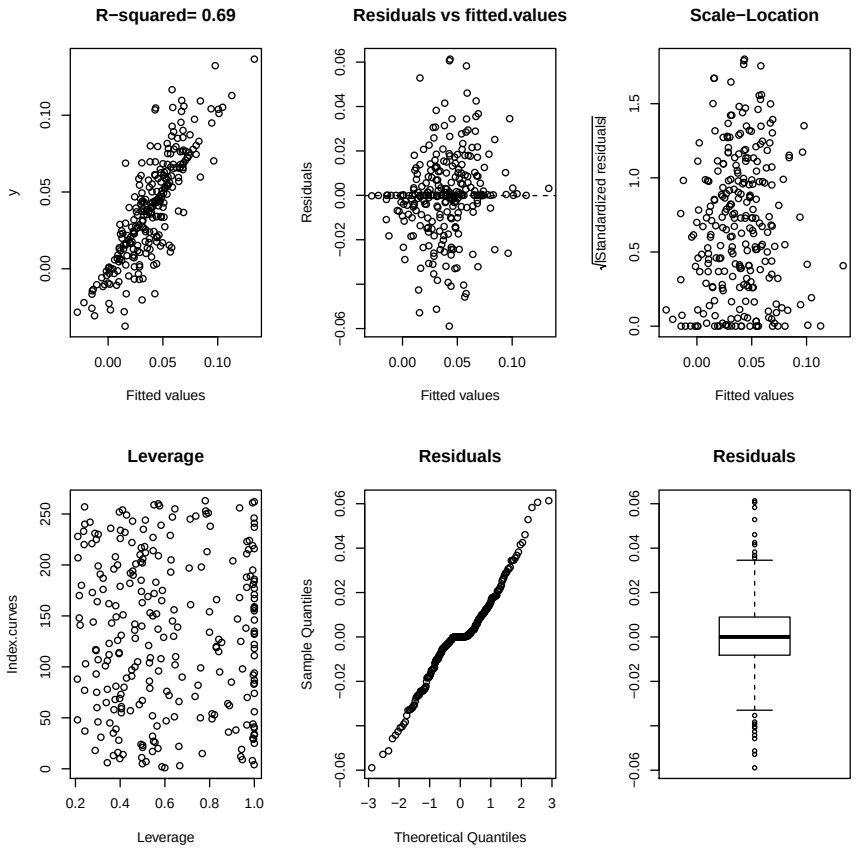


Figure 4.16: Residual and diagnostic plots for functional regression with envelope distance of SIC on SLP.

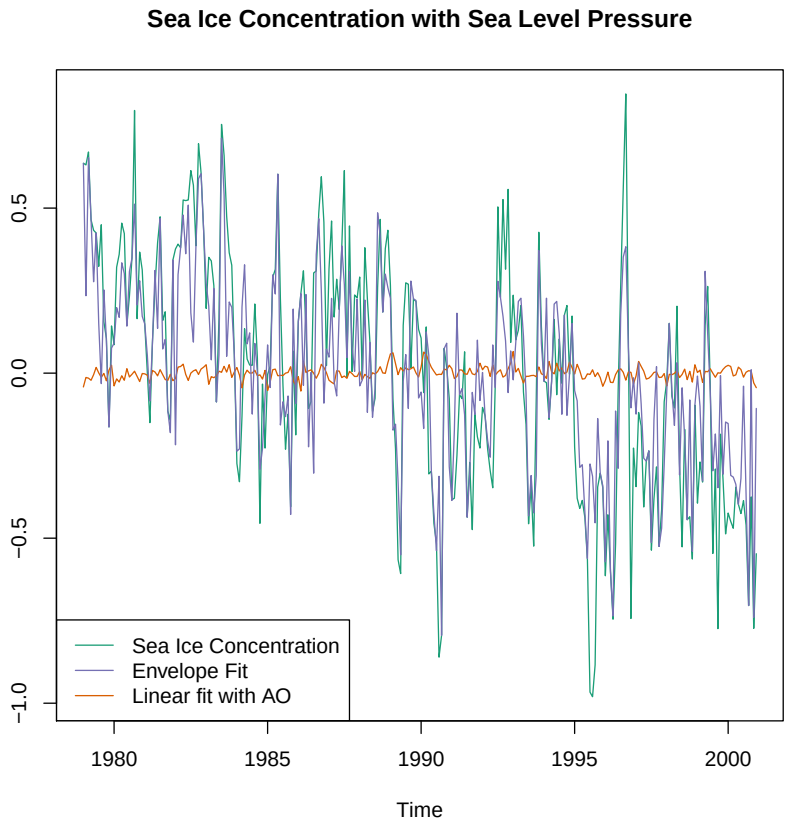


Figure 4.17: Prediction of SIC from SLP with AO and Envelope.

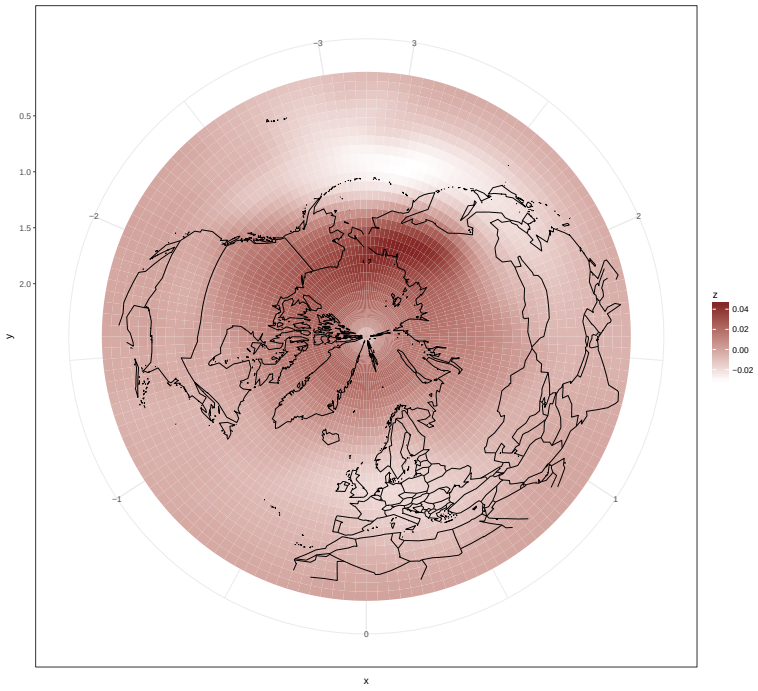


Figure 4.18: Loadings of AO.

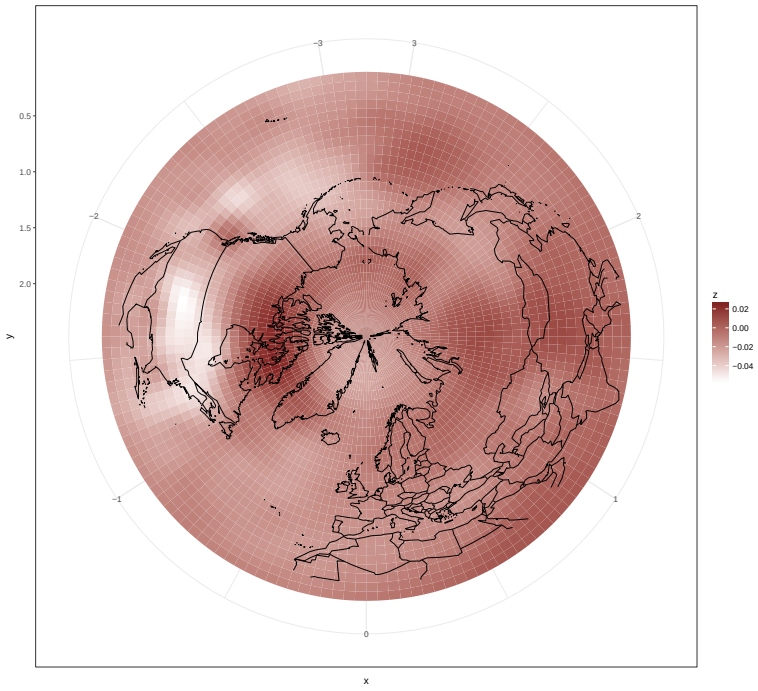


Figure 4.19: Loadings of Envelope.

Bibliography

[1] Mark P. Baldwin, David B. Stephenson, and Ian T. Jolliffe. Spatial Weighting and Iterative Projection Methods for EOFs. *Journal of Climate*, 22(2):234–243, January 2009. ISSN 0894-8755, 1520-0442. doi: 10.1175/2008JCLI2147.1. URL `http://journals.ametsoc.org/doi/abs/10.1175/2008JCLI2147.1`.

[2] O. Banerjee and El Ghaoui. L. and d'Aspremont, A. (2008). *Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res.*, 9:485–516.

[3] Philippe Besse and J. O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311, June 1986. ISSN 0033-3123, 1860-0980. doi: 10.1007/BF02293986. URL `http://link.springer.com/article/10.1007/BF02293986`.

[4] T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106:594–607, 2011.

[5] HervÃl' Cardot. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12(4):503–538, January 2000. ISSN 1048-5252. doi: 10.1080/10485250008832820. URL http://dx.doi.org/10.1080/10485250008832820.

[6] HervÃl' Cardot, FrÃl'dÃl'ric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, pages 571–591, 2003. URL http://www.jstor.org/stable/24307112.

[7] P. E. Castro, W. H. Lawton, and E. A. Sylvestre. Principal Modes of Variation for Processes with Continuous Sample Curves. *Technometrics*, 28(4):329–337, 1986. ISSN 0040-1706. doi: 10.2307/1268982. URL http://www.jstor.org/stable/1268982.

[8] Frank Chambers and Michael Ogle, editors. *Climate change: critical concepts in the environment*. Routledge, London ; New York, 2002. ISBN 978-0-415-27656-6 978-0-415-27657-3 978-0-415-27658-0 978-0-415-27659-7 978-0-415-27660-3.

[9] F. Stuart Chapin III, Robert L Jefferies, James F Reynolds, Gaius R Shaver, Josef Svoboda, and Ellen W Chu. *Arctic Ecosystems in a Changing Climate: an Ecophysiological Perspective*. Elsevier Science, Oxford, 1991. ISBN 978-0-323-13842-0. URL http://www.123library.org/book_details/?id=64948. OCLC: 829460348.

[10] Dong Chen, Peter Hall, and Hans-Georg MÃijller. Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, 39(3):1720–1747, June 2011. ISSN 0090-5364, 2168-8966. doi: 10.1214/11-AOS882. URL http://projecteuclid.org/euclid.aos/1311600281.

[11] Josefino C. Comiso, Claire L. Parkinson, Robert Gersten, and Larry Stock. Accelerated decline in the Arctic sea ice cover. *Geophysical Research Letters*, 35(1), January 2008. ISSN 0094-8276. doi: 10.1029/2007GL031972. URL http://doi.wiley.com/10.1029/2007GL031972.

[12] John B. Conway. *A course in functional analysis*. Number 96 in Graduate texts in mathematics. Springer, New York, 2nd ed edition, 1997. ISBN 978-0-387-97245-9.

[13] R. D. Cook. Class Notes Envelope, 2014. bibtex: cook_class_2014.

[14] R. Dennis Cook and Xin Zhang. Algorithms for Envelope Estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300, January 2016. ISSN 1061-8600. doi: 10.1080/10618600.2015.1029577. URL http://dx.doi.org/10.1080/10618600.2015.1029577.

[15] R. Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960, 2010. URL http://www.jstor.org/stable/24309466.

[16] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to

statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154, March 1982. ISSN 0047-259X. doi: 10.1016/0047-259X(82)90088-4. URL http://www.sciencedirect.com/science/article/pii/0047259X82900884.

[17] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[18] Winston Wei Dou, David Pollard, and Harrison H. Zhou. Estimation in functional regression for general exponential families. *The Annals of Statistics*, 40(5):2421–2451, October 2012. ISSN 0090-5364, 2168-8966. doi: 10.1214/12-AOS1027. URL http://projecteuclid.org/euclid.aos/1359987526.

[19] FrÃl'dÃl'ric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer series in statistics. Springer, New York, 2006. ISBN 978-0-387-30369-7. OCLC: ocm70261207.

[20] Vieu Philippe Ferraty, FrÃl'dÃl'ric. *Nonparametric Functional Data Analysis*. 2006.

[21] J. H. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

[22] Nathan P. Gillett, Francis W. Zwiers, Andrew J. Weaver, and Peter A. Stott. Detection of human influence on sea-level pressure. *Nature*, 422(6929):292–294, March 2003. ISSN 00280836. doi: 10.1038/nature01487. URL http://www.nature.com/doifinder/10.1038/nature01487.

[23] Peter Hall and Joel L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, February

2007. ISSN 0090-5364. doi: 10.1214/009053606000000957. URL http://projecteuclid.org/euclid.aos/1181100181.

[24] Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126, 2006. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00535.x/full.

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[26] Larry D. Hinzman, Neil D. Bettez, W. Robert Bolton, F. Stuart Chapin, Mark B. Dyurgerov, Chris L. Fastie, Brad Griffith, Robert D. Hollister, Allen Hope, Henry P. Huntington, Anne M. Jensen, Gensuo J. Jia, Torre Jorgenson, Douglas L. Kane, David R. Klein, Gary Kofinas, Amanda H. Lynch, Andrea H. Lloyd, A. David McGuire, Frederick E. Nelson, Walter C. Oechel, Thomas E. Osterkamp, Charles H. Racine, Vladimir E. Romanovsky, Robert S. Stone, Douglas A. Stow, Matthew Sturm, Craig E. Tweedie, George L. Vourlitis, Marilyn D. Walker, Donald A. Walker, Patrick J. Webber, Jeffrey M. Welker, Kevin S. Winker, and Kenji Yoshikawa. Evidence and Implications of Recent Climate Change in Northern Alaska and Other Arctic Regions. *Climatic Change*, 72(3):251–298, October 2005. ISSN 0165-0009,

1573-1480. doi: 10.1007/s10584-005-5352-2. URL http://link.springer.com/10.1007/s10584-005-5352-2.

[27] Ivana HorovÃą, Philippe Vieu, and JiÅŹÃ■ Zelinka. OPTIMAL CHOICE OF NONPARAMETRIC ESTIMATES OF A DENSITY AND OF ITS DERIVATIVES. *Statistics & Risk Modeling*, 20(1-4), January 2002. ISSN 2196-7040, 2193-1402. doi: 10.1524/strm.2002.20.14.355. URL http://www.degruyter.com/view/j/strm.2002.20.issue-1-4/strm.2002.20.14.355/strm.2002.20.14.355.xml.

[28] Tailen Hsing and Randall L. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley series in probability and statistics. John Wiley and Sons, Inc, Chichester, West Sussex, 2015. ISBN 978-0-470-01691-6.

[29] C. J. Dalzell J. O. Ramsay. Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):539–572, 1991. ISSN 00359246. URL http://www.jstor.org/stable/2345586.

[30] Gareth M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432, August 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00342. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00342/abstract.

[31] William H. Kruskal. Ordinal Measures of Association. *Journal of the American Statistical Association*, 53(284):pp. 814–861, 1958. ISSN 01621459. URL http://www.jstor.org/stable/2281954.

[32] John E. Kutzbach. Empirical Eigenvectors of Sea-Level Pressure, Surface Temperature and Precipitation Complexes over North America. *Journal of Applied Meteorology*, 6(5):791–802, October 1967. ISSN 0021-8952. doi: 10.1175/1520-0450(1967)006<0791:EEOSLP>2.0.CO;2. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0450%281967%29006%3C0791%3AEEOSLP%3E2.0.CO%3B2.

[33] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012. doi: 10.1214/12-AOS1037. URL http://dx.doi.org/10.1214/12-AOS1037.

[34] Clive Loader. *Local regression and likelihood*. Springer Science & Business Media, 2006. URL http://books.google.com/books?hl=en&lr=&id=NpjeBwAAQBAJ&oi=fnd&pg=PA1&dq=%22and+asymptotic+theory.+Largely,+these+chapters+are+independent+of%22+%22is+necessarily+selective.+I+attempt+to+present+results+that+are%22+%22more+general+approach+to+achieve+the+same%22+&ots=wVafYZoO54&sig=ay4Ekmj0zBkQ2hRgLL-kPojqhRQ.

[35] N. Locantore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, K. L. Cohen, Graciela Boente, Ricardo Fraiman, Babette Brumback, Christophe

Croux, Jianqing Fan, Alois Kneip, John I. Marden, Daniel PeÃśa, Javier Prieto, Jim O. Ramsay, Mariano J. Valderrama, Ana M. Aguilera, N. Locantore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, and K. L. Cohen. Robust principal component analysis for functional data. *Test*, 8 (1):1–73, June 1999. ISSN 1133-0686, 1863-8260. doi: 10.1007/BF02595862. URL http://link.springer.com/article/10.1007/BF02595862.

[36] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34:1436–1462, 2006.

[37] Adam H. Monahan, John C. Fyfe, Maarten H. P. Ambaum, David B. Stephenson, and Gerald R. North. Empirical Orthogonal Functions: The Medium is the Message. *Journal of Climate*, 22(24):6501–6514, December 2009. ISSN 0894-8755. doi: 10.1175/2009JCLI3062.1. URL http://journals.ametsoc.org/doi/abs/10.1175/2009JCLI3062.1.

[38] Jeffrey S. Morris. Functional Regression. *Annual Review of Statistics and Its Application*, 2(1):321–359, April 2015. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-010814-020413. URL http://www.annualreviews.org/doi/10.1146/annurev-statistics-010814-020413.

[39] HANS-GEORG MÃIJLLER. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240, 2005. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2005.00429.x/abstract.

[40] Hans-Georg MÃijller and Ulrich StadtmÃijller. Generalized functional linear models. *The Annals of Statistics*, 33(2):774–805, April 2005. ISSN 0090-5364. doi: 10.1214/009053604000001156. URL http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1117114336/.

[41] Gerald R. North, Thomas L. Bell, Robert F. Cahalan, and Fanthune J. Moeng. Sampling Errors in the Estimation of Empirical Orthogonal Functions. *Monthly Weather Review*, 110(7):699–706, July 1982. ISSN 0027-0644. doi: 10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(1982)110%3C0699%3ASEITEO%3E2.0.CO%3B2.

[42] Scott B. Power and Lawrence A. Mysak. On the interannual variability of arctic seaâĂŘlevel pressure and sea ice. *Atmosphere-Ocean*, 30(4):551–577, December 1992. ISSN 0705-5900, 1480-9214. doi: 10.1080/07055900.1992.9649455. URL http://www.tandfonline.com/doi/abs/10.1080/07055900.1992.9649455.

[43] James Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Springer New York, New York, NY, 2009. ISBN 978-0-387-98184-0 978-0-387-98185-7. URL http://link.springer.com/10.1007/978-0-387-98185-7.

[44] Ramsey J and Silverman B W. *Functional Data Analysis*. 2005.

[45] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. *Model selection*

in Gaussian graphical models: High-dimensional consistency of l1-regularized MLE. In Advances in Neural Information Processing Systems 22. MIT Press, Cambridge, MA, 2009.

[46] John A. Rice. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, pages 631–647, 2004. URL `http://www.jstor.org/stable/24307409`.

[47] Ignatius G. Rigor, John M. Wallace, and Roger L. Colony. Response of sea ice to the Arctic Oscillation. *Journal of Climate*, 15(18): 2648–2663, 2002. URL `http://journals.ametsoc.org/doi/abs/10.1175/1520-0442(2002)015%3C2648%3AROSITT%3E2.0.CO%3B2`.

[48] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008.

[49] Z. Shi. Small ball probabilities for a Wiener process under weighted sup-norms, with an application to the supremum of bessel local times. *Journal of Theoretical Probability*, 9(4):915–929, October 1996. ISSN 0894-9840, 1572-9230. doi: 10.1007/BF02214257. URL `http://link.springer.com/article/10.1007/BF02214257`.

[50] David W. J. Thompson and John M. Wallace. The Arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical Research Letters*, 25(9):1297–1300, May 1998. ISSN 00948276. doi: 10.1029/98GL00950. URL `http://doi.wiley.com/10.1029/98GL00950`.

[51] John E. Walsh, William L. Chapman, and Timothy L. Shy. Recent Decrease of Sea Level Pressure in the Central Arctic. *Journal of Climate*, 9(2):480–486, February 1996. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(1996)009<0480:RDOSLP>2.0.CO;2. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0442%281996%29009%3C0480%3ARDOSLP%3E2.0.CO%3B2.

[52] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg MÃijller. Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, June 2016. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-041715-033624. URL http://www.annualreviews.org/doi/10.1146/annurev-statistics-041715-033624.

[53] Suojin Wang, Lianfen Qian, and Raymond J. Carroll. Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika*, 97(1):79–93, 2010. URL http://biomet.oxfordjournals.org/content/97/1/79.short.

[54] Hulin Wu and Jin-Ting Zhang. *Nonparametric regression methods for longitudinal data analysis: [mixed-effects modeling approaches]*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J, 2006. ISBN 978-0-471-48350-2. OCLC: ocm62525265.

[55] Luo Xiao, Yingxing Li, and David Ruppert. Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series*

B (Statistical Methodology), 75(3):577–599, 2013. ISSN 1467-9868. doi: 10.1111/rssb.12007. URL http://dx.doi.org/10.1111/rssb.12007.

[56] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286, 2010.

[57] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.

[58] Xiaoke Zhang and Jane-Ling Wang. From Sparse to Dense Functional Data and Beyond. *Annals of Statistics (Submitted)*.

[59] Xin Zhao, James Stephen Marron, and Martin T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, pages 789–808, 2004. URL http://www.jstor.org/stable/24307416.

Appendices

Appendix A

Computation for Multiple Testing

A.1 R package mhtboot

Our method for multiple hypothesis testing using bootstrap distribution of the p values has been implemented in an R package 'mhtboot' and is available through CRAN. Here we provide a short description of the implementation and other functions provided by this package.

We know that under null hypothesis p value is an uniform$(0, 1)$ random variable. If the hypothesis are independent, then so are the p values. We use this intuition to construct a statistic that can be used to test multiple hypothesis.

We consider the case of coordinate wise testing of a data vector. If the data matrix X is of dimension $I \times J$ then without loss of generality, we consider the following set of hypothesis for $j = 1, 2, \cdots, J$.

$$H_0^{(j)} : x_j = 0$$
$$H_1^{(j)} : x_j \neq 0$$

Where x_j denotes the jth coordinate of X. This set of tests can be performed in a variety of ways, for the purpose of illustration, we choose to use one sample t-tests.

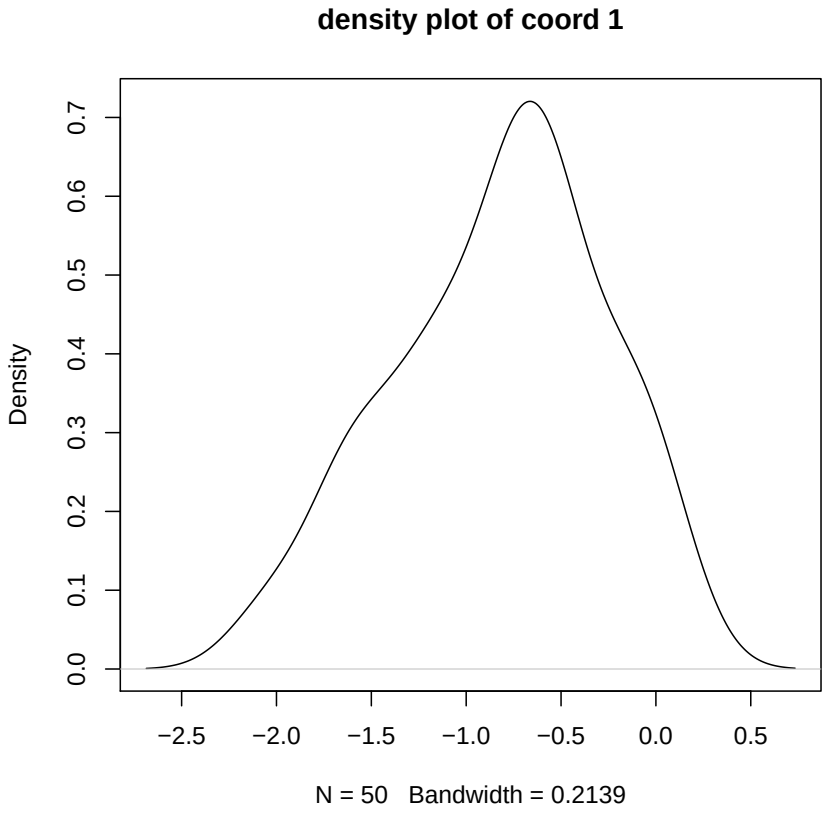
The general procedure for testing multiple hypothesis is testing independently then use a correction method to correct for family wise error rate. Here we are not using any such correction, as we are using a bootstrap based approach. We first generate bootstrap samples of our data, let us denote the bootstrap samples by $X^{\{b\}}$ for $b \in \{1, 2, \cdots, B\}$, where B is the bootstrap sample size. For each of these samples we can perform J tests in parallel and collect the p values. We denote

the p values from jth coordinate of bth bootstrap sample by $p_j^{\{b\}}$. We then use a monotone transformation of the p values for better visualizing. The transformed order statistics are collected as shown.

$$
\begin{aligned}
Z_{(0)}^{\{b\}} &\stackrel{\text{def}}{=} 0 \\
Z_{(j)}^{\{b\}} &= -\log(1 - p_j^{\{b\}}) \\
W_j^{\{b\}} &\stackrel{\text{def}}{=} Z_{(j)}^{\{b\}} - Z_{(j-1)}^{\{b\}} - (n + 1 - j)^{-1}
\end{aligned}
$$

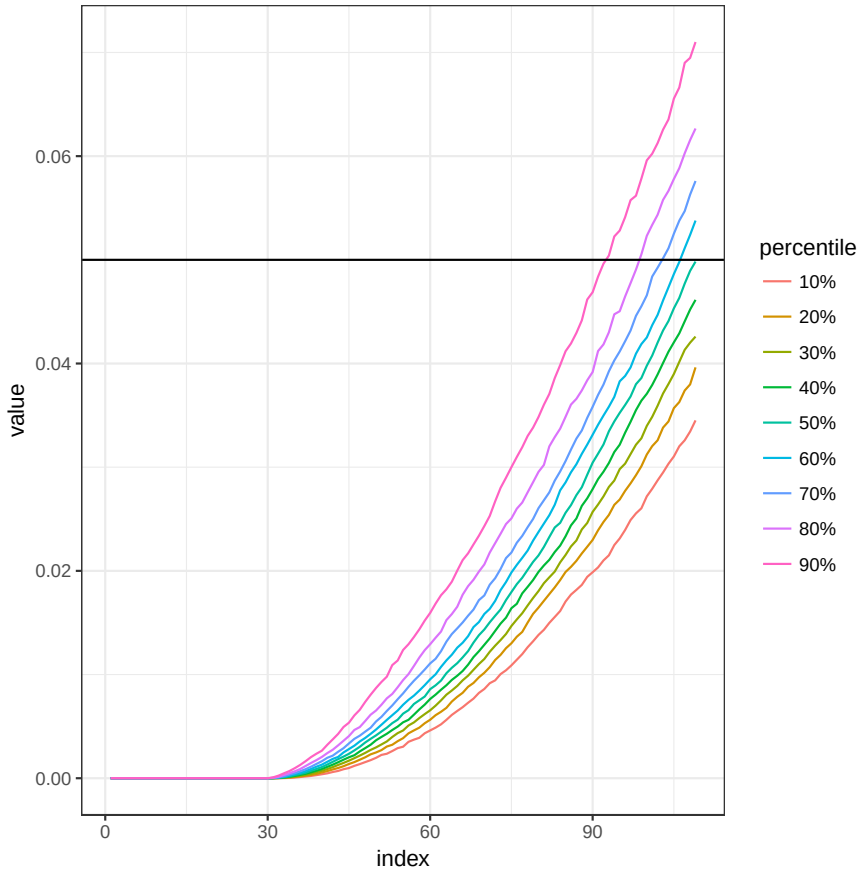
We demonstrate the proceedure below with a simulated example. The function datgen is used to simulate data from a multivariate normal distribution.

```
> set.seed(12345)
> suppressMessages(library(mhtboot))
> n = 50;m = 500;m0 = 30;
> sigeff = 1;
> Sigma <- 0.25*diag(m)
> X <- datgen(n,m,m0,sigeff,Sigma = Sigma)
> plot(density(X[,1]),main="density plot of coord 1")
```



We then generate the distribution of the p values using bootstrap. This is impleemnted in the function series pboot. Here we are using the one sample version of the function. The plotpboot function is used to generate the quantile plot of the distribution of the p values.

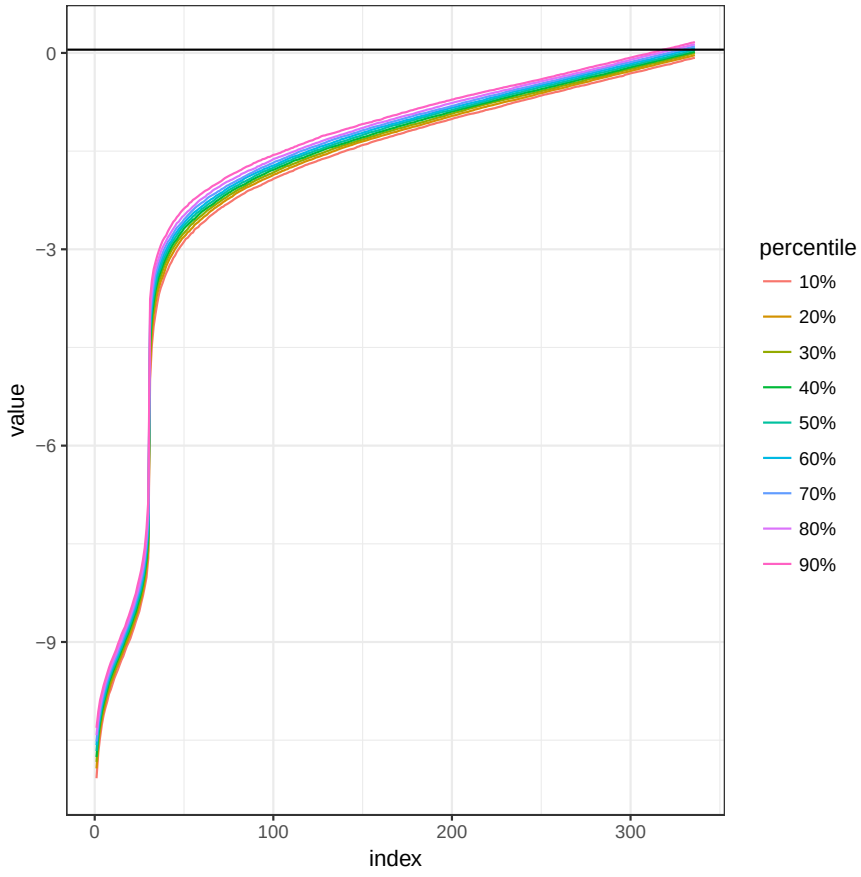
```
> porder <- pboot.1sample(X=X,B=500,ncpus = 16)
> plotpboot(porder = porder)
```



This approach can be extended to any set of hypothesis, in the package we also provide a function for two sample tests. Both of these functions can be used for user given tests as they accept test statistic as a parameter.

We can transform the order statistics of the p values using a monotone transformation. We show here the transformation using inverse normal cdf function.

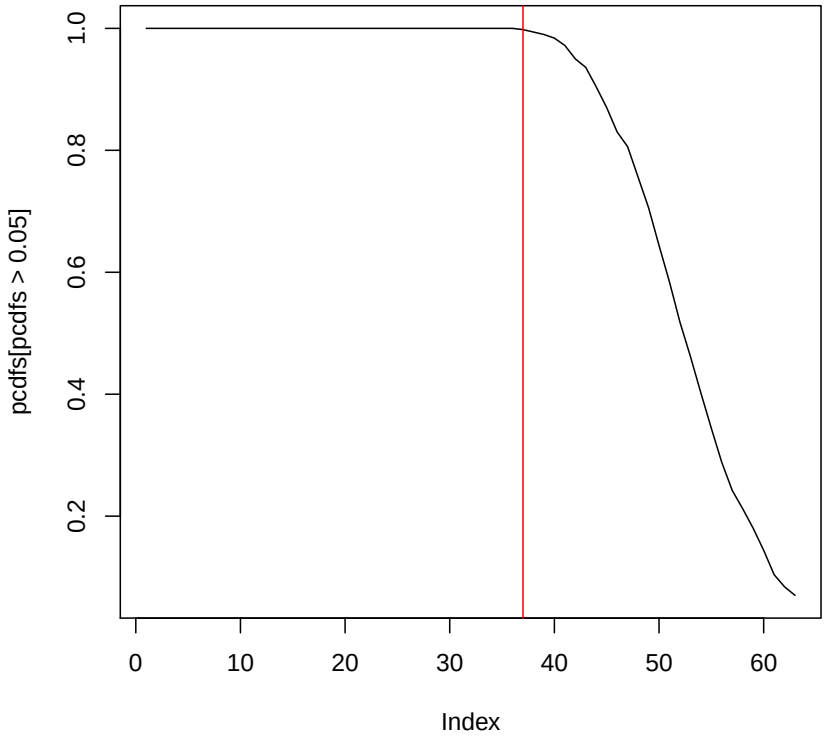
```
> porder.tr <- ptrans(porder,trans="normal")
> plotpboot(porder.tr)
```



We can also look at the points where each coordinates hit a certain probability.

```
> porder <- ptrans(porder = porder)
> hitplots(porder = porder,alpha = 0.005)

[1] 37
```



Once we have the distribution of the p values, we can use them to detect the change point in their distribution. This is done through qelbow function.

```
> out <- qelbow(porder = porder)
> out

dav dlm
 42  34
```