Scalable Estimation and Testing for Complex, High-Dimensional Data

Ruijin Lu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Hongxiao Zhu, Chair
Inyoung Kim
Xinwei Deng
Xiaowei Wu

July 25, 2019
Blacksburg, Virginia

Keywords: Functional data testing, randomization method, basis decomposition, approximate Bayesian computation (ABC), wavelet decomposition, Gaussian Process surrogate model, fluctuation analysis, mutation probability estimation, birth-death process model, region detection, fluorescence spectroscopy data, Alzheimer's disease Initiative, macaque monkeys, spherical wavelet, cortical early development, hemispheric asymmetries, sparse longitudinal data analysis, cortical thickness, data supported on manifolds

Scalable Estimation and Testing for Complex, High-Dimensional Data

Ruijin Lu

(ABSTRACT)

With modern high-throughput technologies, scientists can now collect high-dimensional data of various forms, including brain images, medical spectrum curves, engineering signals, etc. These data provide a rich source of information on disease development, cell evolvement, engineering systems, and many other scientific phenomena. To achieve a clearer understanding of the underlying mechanism, one needs a fast and reliable analytical approach to extract useful information from the wealth of data.

The goal of this dissertation is to develop novel methods that enable scalable estimation, testing, and analysis of complex, high-dimensional data. It contains three parts: parameter estimation based on complex data, powerful testing of functional data, and the analysis of functional data supported on manifolds. The first part focuses on a family of parameter estimation problems in which the relationship between data and the underlying parameters cannot be explicitly specified using a likelihood function. We introduce a wavelet-based approximate Bayesian computation approach that is likelihood-free and computationally scalable. This approach will be applied to two applications: estimating mutation rates of a generalized birth-death process based on fluctuation experimental data and estimating the parameters of targets based on foliage echoes.

The second part focuses on functional testing. We consider using multiple testing in basis-space via p-value guided compression. Our theoretical results demonstrate that, under regularity conditions, the Westfall-Young randomization test in basis space achieves strong control of family-wise error rate and asymptotic optimality, and furthermore, appropriate compression in basis space leads to improved power as compared to point-wise testing in data domain or basis-space testing without compression. The effectiveness of the proposed procedure is demonstrated through two applications: the detection of regions of spectral curves associated with pre-cancer using 1-dimensional fluorescence spectroscopy data and the detection of disease-related regions using 3-dimensional Alzheimer's Disease neuroimaging data.

The third part focuses on analyzing data measured on the cortical surfaces of monkeys' brains during their early development, and subjects are measured on misaligned time markers. In this analysis, we examine the asymmetric patterns and increase/decrease trend in the monkeys' brains across time.

Scalable Estimation and Testing for Complex, High-Dimensional Data

Ruijin Lu

(GENERAL AUDIENCE ABSTRACT)

With modern high-throughput technologies, scientists can now collect high-dimensional data of various forms, including brain images, medical spectrum curves, engineering signals, and biological measurements. These data provide a rich source of information on disease development, engineering systems, and many other scientific phenomena.

The goal of this dissertation is to develop novel methods that enable scalable estimation, testing, and analysis of complex, high-dimensional data. It contains three parts: parameter estimation based on complex biological and engineering data, powerful testing of high-dimensional functional data, and the analysis of functional data supported on manifolds.

The first part focuses on a family of parameter estimation problems in which the relationship between data and the underlying parameters cannot be explicitly specified using a likelihood function. We introduce a computation-based statistical approach that achieves efficient parameter estimation scalable to high-dimensional functional data.

The second part focuses on developing a powerful testing method for functional data that can be used to detect important regions. We will show nice properties of our approach. The effectiveness of this testing approach will be demonstrated using two applications: the detection of regions of the spectrum that are related to pre-cancer using fluorescence spectroscopy data and the detection of disease-related regions using brain image data.

The third part focuses on analyzing brain cortical thickness data, measured on the cortical surfaces of monkeys' brains during early development. Subjects are measured on misaligned time-markers. By using functional data estimation and testing approach, we are able to: (1) identify asymmetric regions between their right and left brains across time, and (2) identify spatial regions on the cortical surface that reflect increase or decrease in cortical measurements over time.

Dedication

To my dearest family.

Acknowledgments

This dissertation would not have been possible without the help and support of my advisor, committee, collaborators, family, and friends. I would first like to express my deepest gratitude to my advisor, Dr. Hongxiao Zhu, who has inspired and guided my research work. She is passionate and full of inspiration. With her guidance, I have the opportunity to touch many exciting and challenging areas of statistical research, which motivates me to explore more in the future. During the last four years of academic cooperation, she gave me persistent help and encouragement in overcoming difficulties and improving my academic independence as well as communication skills. I feel so grateful to have the chance of working with her.

I want to thank my committee members, Dr. Inyoung Kim, Dr. Xinwei Deng, and Dr. Xiaowei Wu. Their valuable advice and thought-provoking questions have broadened my view and pushed me to think deeper. I would also like to thank the statistics department of Virginia Tech for providing various courses on cutting-edge topics and plenty of opportunities in teaching and collaborating. I would also like to thank Sarah Boudreau and Dana Omirova from the writing center for their kind help in proof-reading my dissertation.

I would like to thank my collaborators, who have generously shared their valuable and meaningful data with us. Their data have motivated this dissertation. Thanks to Dr. Dennis D. Cox from the statistics department of Rice University for sharing cervical pre-cancer data with us. Thanks to Dr. Gang Li from the school of medicine at the University of North Carolina for sharing data of cortical measurements of monkeys' brains with us. Thanks to Dr. Rolf Mueller and Dr. Chen Ming from the mechanical engineering department of Virginia Tech for sharing sonar simulation code with us.

I hope to thank my family, in particular, my mother, Juping Liu, my father, Min Lu, and

my brother, Junshi Liu for their endless and unconditional love and support. I can always recall the scene that, twenty-four years ago, on my first day of school, my mother waved goodbye to me at the school gate. Since then, they keep standing by me and encouraging me to pursue what I like. Their strong faith in life gives me the strength to face any difficulties. It is my honor to have them as my family.

Last, I hope to thank all my friends. Their friendship has a magical power to keep me positive. Special thanks to my best friend and roommate, Boya Zhang, for her willingness to share my happiness and sorrow.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Overview

Complex, high-dimensional data like multi-platform genomics and brain imaging data can be used to test scientific hypotheses, understand the functionality of biological systems, or discover biomarkers that provide insights into disease diagnosis and treatment. These data, ranging in size from gigabytes to terabytes, represent the leading edge of the onslaught of Big Data. However, as the richness of high-dimensional data continues to grow, some critical questions remain unanswered, such as how to uncover the underlying systematic patterns by filtering out noise and how to extract important features for assessment and decision-making. These questions point towards a challenge that statisticians and data scientists must face — inferential tools must be methodically efficient, theoretically sound, and computationally scalable to the size and dimension of Big Data so that an ideal level of inferential accuracy can be achieved within a limited time budget.

The primary challenges in theory and methods lie in several aspects. First, the high dimensionality makes traditional univariate or multivariate analysis intractable. Second, the high resolution and high correlation between measurement points distinguish some high-dimensional data from traditional data formats, constituting a new data type called functional data—realizations of random functions varying over a continuum—that requires new theoretical/methodological development for its analysis. Third, the complex data structures,

including spatial/temporal correlations, and correlation induced by hierarchical experimental designs, requires more delicate modeling assumptions than independence. Finally, some data, such as the brain cortical surface data, involve geometric structures which need special theoretical and methodological care.

In addition to theoretical and methodological challenges, another important concern is its computational scalability. Despite recent developments, most existing statistical procedures fail to scale to high-dimensionality and result in runtimes that render them unusable on large-scale datasets. Faced with this situation, ad hoc procedures are often adopted which either ignore important structures in the data or rely on subjectively selected features which could miss important information in the data. Such ad hoc procedures perhaps provide algorithmic guarantees but may not provide statistical guarantees and hence could lead to biased conclusions.

In this dissertation, I will focus on developing a suite of theoretically rigorous and computationally scalable statistical methods for parameter estimation and hypothesis testing for various types of complex high-dimensional data. In particular, three types of problems will be considered:

- **Estimating Parameters in Complex Systems** We consider a family of parameter estimation problems involving complex physical systems and/or high-dimensional data. In these problems, the relationship between data and the underlying parameters cannot be explicitly specified using a likelihood function. These situations often occur when data arises from a complex system and only numerical simulations can be used to describe the underlying data-generating mechanism. We will develop a scalable likelihood-free and simulation-based approach to estimate parameters.

- **Hypothesis Testing of High-Dimensional Functional Data** Pointwise testing

suffers from power loss in the high-dimensional setting due to the high correlation between measurement points. We consider a randomization-based multiple testing procedure, the Westfall-Young Randomization Tests in basis-space via lossless or near-lossless compression, to improve the power of testing and to identify significant regions on functional data. Theoretical properties will be investigated, such as the control of family-wise error rate, the power improvement under appropriate truncation, and the asymptotic optimality.

- **Analysis of Longitudinal Functional Data Supported on Manifolds** Motivated by monkeys' cortical measurements data, we consider the analysis of longitudinal functional data that are spatially dense, longitudinally sparse, irregularly collected and supported on manifolds. The goal is to identify cortical regions of the monkeys' brains that are significantly asymmetric or developed across the early development stages of monkeys, from 0 to 36 months. We adopt efficient compression using spherical wavelets and perform component-wise smoothing and curve fitting in the spherical wavelet domain. Randomization-based multiple testing procedures will be performed to identify significant wavelet components. Results will be transformed and visualized in the original data domain.

Compared with existing methods, methods proposed in this dissertation enable scalable inference through parsimonious data representation, compression, and parallel computation. For example, by combining sparse multi-resolution representation such as wavelet transformation with near-lossless compression, we can reduce the dimension of a functional observation from $O(10^5)$ to $O(10^3)$. Most computation is carried out in the compressed domain in an "embarrassingly parallel fashion," so that computation units work independently with no or very little communication between the components. Furthermore, the inferential results are often transformed back to the original data space, enabling straightforward interpretation

and visualization. Such designs guarantee flexibility, automation, and statistically rigorousness, thus facilitating effectiveness of knowledge discovery from complex, high-dimensional data.

1.2 Motivation Examples

The methods proposed in this dissertation are primarily motivated by data arising from several real applications. This section demonstrates five motivating examples of complex systems or data with complex structures in the field of computational biology, biophysical engineering, biomedical sciences and neuroscience. The corresponding methodology for analyzing such data and conducting statistical inference will be investigated in great detail in Chapters 2 - 4.

1.2.1 Estimating mutation parameters in a generalized birth-death process

The parameter estimation problem involves data arising from a complex system. Due to the complexity of the system, traditional estimation approaches become infeasible or inappropriate. However, the data-generating mechanism underlying such a system can usually be described by numerical simulations, which provides a possible solution to the parameter estimation problem. Our first motivating example concerns data generated from a cell growth and mutation process, which, without loss of generality, can be treated as a "generalized" birth-death process.

Consider a biological process of cell growth and mutation, in which a non-mutant cell grows to a cell population according to a certain branching rule (in most cases, Markov branching

is assumed, and cell death is ignored), and mutations occur randomly at the time of cell division with backward mutations not allowed. Figure 1.1 below shows an example of such a process (a binary fission Markov branching process) with two different schemes, pre- and post-division mutation. It can be seen that, under certain model assumptions, the non-mutant counts can be modeled by a birth-death process.

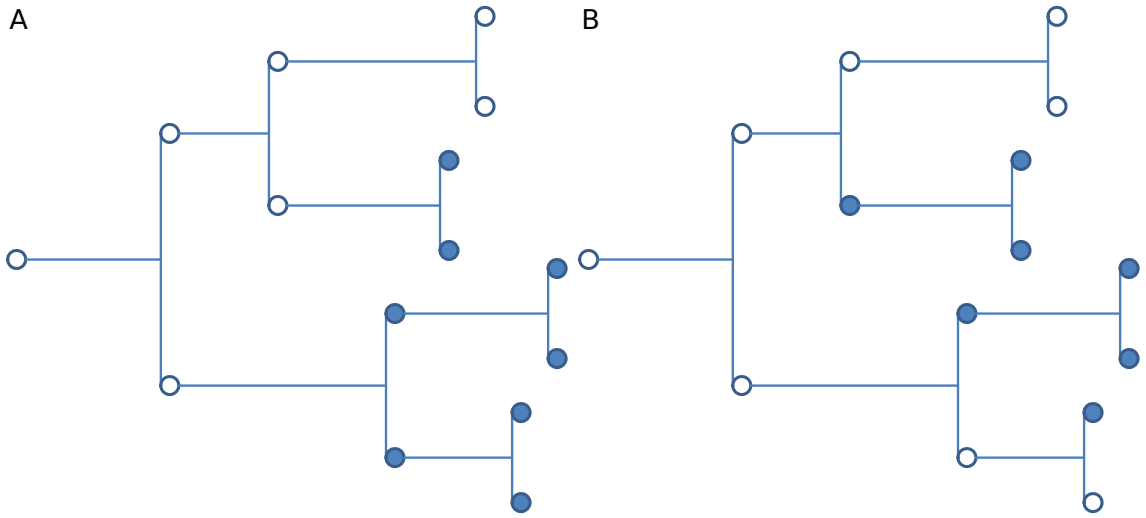

Figure 1.1: An illustration of the cell growth and mutation process (a binary fission Markov branching process); empty and filled circles represent non-mutant and mutant cells, respectively. A: Pre-division mutation; B: Post-division mutation.

The key problem in this process is to estimate the mutation rate based on the observed data (usually mutant and non-mutant counts) at a given time point. In practice, the real bioassay data are usually collected under dynamic experimental conditions, which, in turn, increases the complexity of modeling and hence the difficulty of mutation rate estimation. For example, we may need to consider the following relaxed assumptions for the model: the non-Markovian branching (i.e., the cell life spans are not i.i.d. exponential), the differential growth of mutant and non-mutant cells, and most importantly, the non-constant mutation rate over time. Conceptually, we may call such a flexible model with only the "birth and death" property retained the "generalized birth-death process". Given the complexity of the

parameters that govern the generalized birth-death process, it would be difficult to estimate the mutation parameters via traditional inference procedure.

However, it should be noted that the simulation procedure of such a complex process is quite straightforward. Giving the setting of different parameters, e.g., the life span distribution and the offspring distribution of the mutant and non-mutant cells, and the mutation rate trajectory over time, it is easy to simulate the sample path of the mutant and non-mutant counts. Therefore, the essential question upon this point is: Can we estimate the dynamic mutation rate of a real cell growth and mutation process by borrowing strength from large-scale, repeated simulations?

1.2.2 Estimating parameters in a sonar simulation system

Another example for the parameter estimation problem is the estimation of parameters in a sonar simulation system. We consider foliage-echo data arising from a sonar study. During the data collection, an active sonar system is used, whose working mechanism is demonstrated in Figure 1.2. Specifically, the active sonar system consists of an emitter that ensonifies the environment and a receiver that records the returning echoes. The transmitter emits acoustic waves and the receiver collects echoes reflected from objects in the environment. The echo signals carry information about the targets, hence they have been used for various identification and navigation tasks [65]. In natural environments, an echo signal is the superposition of reflected waveforms from numerous scatterers, such as foliage leaves and rocks in uneven natural terrains, thus it is highly stochastic.

While the mechanism of sound propagation and reflection is complicated, we are able to simulate foliage echoes using a simulator by applying acoustic laws under simplified assumptions. Specifically, we have established a computational model to simulate a natural sonar

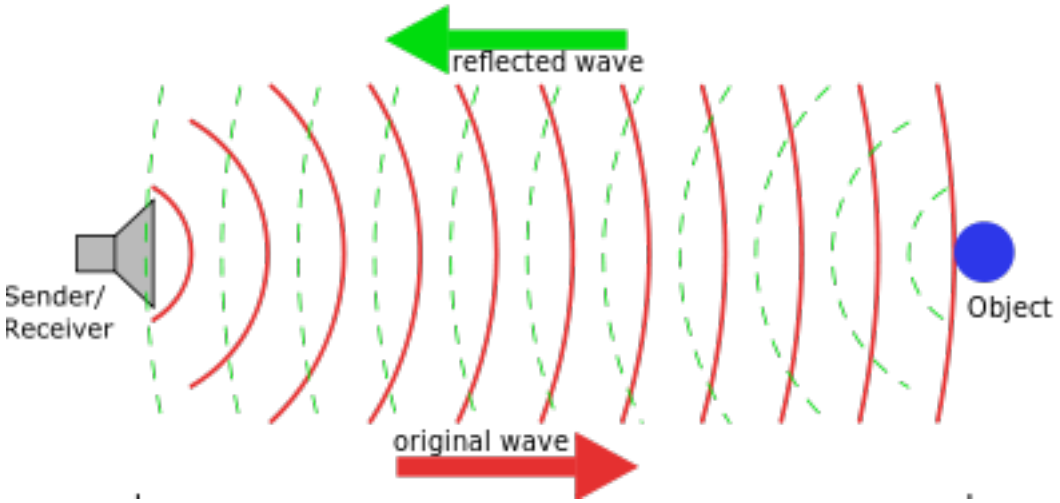

Figure 1.2: The principle of an active sonar. This figure was created based on an online figure available at Wikipedia [72](https://en.wikipedia.org/wiki/Sonar)

scene in a three-dimensional (3-d) space. The scene is demonstrated in Figure 1.3 (a), which consists of an active sonar sensor and a cluster of tree leaves. The sensor was located at the origin. It emitted ultrasonic waves towards the positive x-axis direction.

The tree foliages were uniformly located in a $[1, 10] \times [-2, 2] \times [-2, 2]$ region in 3-d. The total number of leaves was determined by the leaf density—the counts of leaves per cubic meter, denoted by θ_1. The leaf shapes were approximated by disks with radius randomly sampled from a normal distribution $N(\theta_2, 0.1\theta_2)$, where θ_2 denotes the mean radius. The orientation of each leaf relative to the sonar was determined by the angle (in degrees) between the leaf's normal direction and the sonar-leaf center line. These angles were randomly simulated from a truncated normal distribution $N(x \mid \theta_3, 5)1_{\{0<x<90\}}$, where θ_3 denotes the mean angle. With these setups and a pre-specification of the acoustic properties of the sonar sensor, echoes were simulated following acoustic laws for sound reflections [5].

The above simulation model constitutes a physical system with three inputs: leaf intensity (θ_1), mean leaf radius (θ_2), and mean leaf orientation (θ_3). The output is an echo signal as demonstrated in Figure 1.3 (b). The output echo signal is a temporal waveform measured

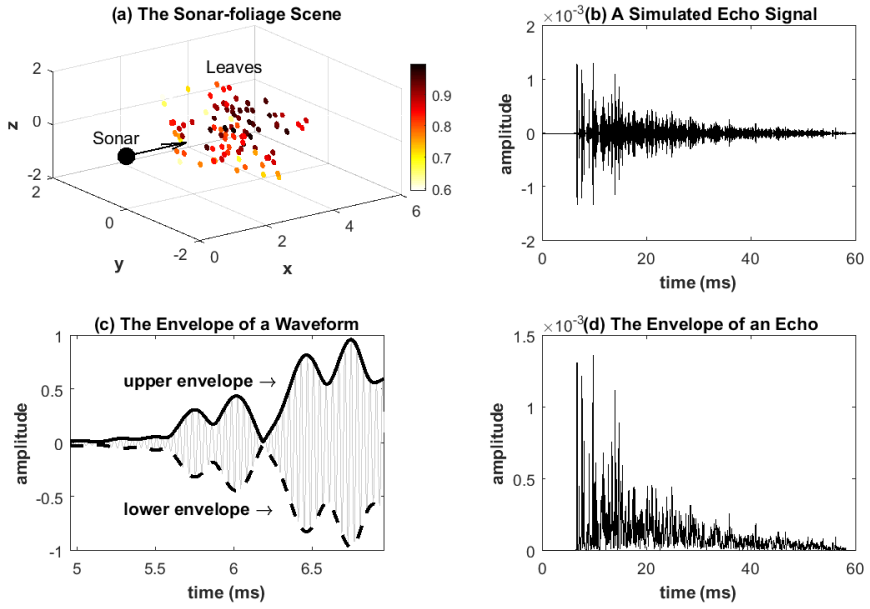

Figure 1.3: The foliage-echo simulation. (a) The sonar scene in 3-d. The sonar sensor is located at the origin. It emits ultrasonic waves towards the positive x-axis direction. Leaves are uniformly placed in a $[1, 10] \times [-2, 2] \times [-2, 2]$ region in 3-d. The color indicates the amount of the sound the leaves receive/reflect (scaled to $[0, 1]$). These quantities are calculated based on the acoustic properties of the sonar. (b) A simulated echo signal with leaf density of 30 (number of leaves per cubic meter), leaf radius of 0.0171 (in meter), and leaf orientation of 45 (in degree, the angle between the leaf normal direction and the sonar-leaf center line). (c) A demonstration of the upper and lower envelopes of an waveform. (d) The echo envelope extracted from the echo signal in (b).

from 0 to 60 milliseconds with a sampling rate of 400 kHz. The total number of measurement points is $24,000$ for each echo. The parameters $(\theta_1, \theta_2, \theta_3)$ summarize the statistical properties of the foliage targets. Therefore, estimating these parameters based on the echo signals provides us with knowledge of the targets.

The foliage-echo data represent a general class of physical systems with functional data outputs. Our goal of the study is to estimate the parameters of the tree foliages (i.e., the density, size, and orientation) based on the echo signals that are either captured by the sonar device or simulated by the computational model.

Directly modeling the echo signals is difficult because the echoes contain redundant information from the emitted waves (the "carrier" waves). We therefore perform a preprocessing step by extracting the *envelopes* of the echo signals. The envelope of a signal is the boundary curve within which all amplitude values of the signal are contained. A conceptual demonstration is shown in Figure 1.3 (c). The envelope of an echo retains the target-specific information through capturing the lower frequency amplitude variations, which makes it an ideal representation of the echo signal. In the sonar echo data, since the upper and the lower envelopes are always symmetric, we only consider the upper envelopes in our data analysis. The envelope signal extracted from the echo in Figure 1.3 (b) is shown in Figure 1.3 (d).

1.2.3 Testing based on fluorescence spectroscopy data

Another family of problems we will investigate is the hypothesis testing of high-dimensional data in functional form. One motivating example is the fluorescence spectroscopy data in a cervical pre-cancer study. Fluorescence spectroscopy is a technique which captures the spectra of fluorescent lights emitted by a given material. It provides doctors with a non-invasive, low-cost alternative to the existing approaches for the diagnosis and assessment of early stage cervical cancer. The data we look at in this paper are collected from a clinical study using multiple fluorescence spectra to detect cervical abnormalities.

Each measurement consists of multiple spectral curves measured on the same cervical tissue site according to the following procedure. First, we illuminate the cervical tissue site with an excitation light at a certain fixed wavelength. The excitation light is absorbed by various endogenous fluorescent molecules in the tissue, resulting in the emission of fluorescent light. The emitted fluorescent light is then captured by an optical detector which produces a spectrum in the shape of a smooth curve. We repeat the above procedure by varying the

wavelength of the excitation light at a sequence of excitation wavelengths, which gives multiple spectral curves for each measurement. An example measurement is shown in Figure 1.4. It contains 16 spectral curves with different excitation wavelengths (ranging from 330 nm to 480 nm with step size 10 nm). At each excitation wavelength, there is a smooth spectral curve that contains the intensity measurements on an interval of emission wavelengths ranging from 385 nm to 700 nm. Here, we have normalized the fluorescence intensities through dividing them by the excitation light energy measurement of each excitation wavelength, so that they are comparable across different excitation wavelengths. The spectral pattern is usually shown by a color band, representing the intensities of each spectral curve according to different combinations of excitation wavelength and emission wavelength, as shown in Figure 1.4(b). We refer to this kind of data as the excitation-emission matrices (EEMs).

The EEM data studied in this dissertation were collected at the same clinic (British Columbia Cancer Agency, Vancouver, CA) with the same instrument (called FastEEM3). The preprocessing of this data followed a six-step procedure, whose details were described by Marín et al. [39]. The dataset contains 534 EEM measurements of 143 pre-cancer samples and 391 normal samples. The pre-cancer samples include tissue sites diagnosed as cervical intraepithelial neoplasia (CIN) II or worse, and normal samples refering to those diagnosed as CIN I or better. All measurements came from sites with colposcopic tissue type “squamous”, and from pre-menopausal patients.

The goal of our study is to identify systematic differences between pre-cancer and normal samples based on the EEM measurements. The testing result will help us flag regions on EEM measurements that reflect differences between the pre-cancer samples and the normal ones. As shown in Figure 1.4, the spectral curves are correlated because of the natural ordering of the excitation wavelengths. Thus, we need a testing approach that can naturally incorporate intra- and inter-correlation of the curves, has enough power to detect significant

regions, and is computationally scalable to the dimensionality.

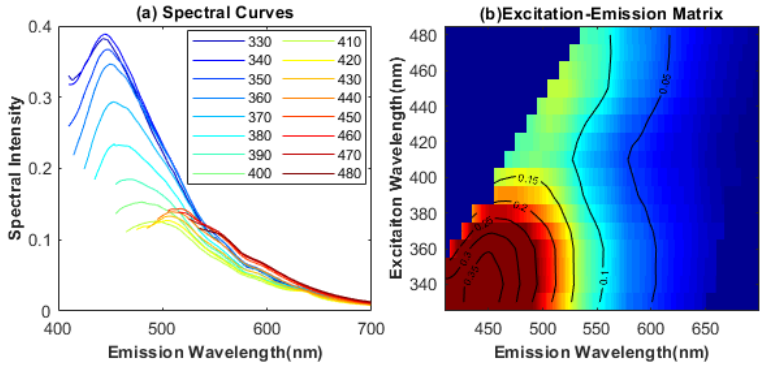

Figure 1.4: Plots of one fluorescence spectroscopy measurement: (a) Fluorescence spectral curves at different excitation wavelengths in nanometers(nms), (b) Image plot of the excitation-emission matrix (EEM) with color standing for intensity.

1.2.4 Testing based on tensor-based morphometry images of human brain

The second motivating example for the high-dimensional testing problem is a 3-d brain image dataset obtained from the NIH Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://www.loni.ucla.edu/ADNI). ADNI is an ongoing open study to develop clinical imaging, genetic, and biochemical bio-markers for the early detection and tracking of Alzheimer's disease (AD). Participants are followed and reassessed over time to track the pathology of the disease as it progresses.

The data is publicly downloadable, including MR scans of 188 AD patients, 385 individuals with mild cognition impairment (MCI), and 228 control elderly. The data is preprocessed using tensor-based morphometry (TBM), an image analysis technique that measures brain volumetric differences relative to a common anatomical template. TBM produces 3-D brain images that characterize brain atrophy corresponding to tissue loss. Details of the prepro-

cessing steps are described in Hua and Thompson [23]. In Figure 1.5, we plot the point-wise differences between the samples means for AD-NL and MCI-NL groups at a fixed slice (z=114).

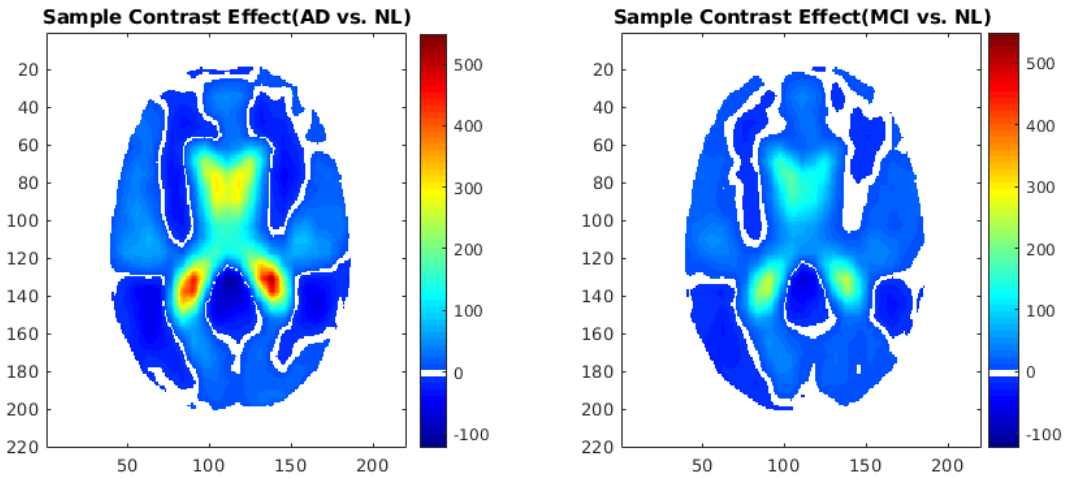

Figure 1.5: Sample mean contrast of MRI signal intensity between AD and Normal, and MCI and Normal. Left: mean contrast between AD and normal groups; Right: mean contrast between MCI and normal groups.

Our goal in this study is to use hypothesis testing to detect regions that reflect volumetric brain differences between AD vs. NL and between MCI vs. NL. This study represents high-dimensional functional data that are challenging to perform multiple testing on. Because of the high dimensionality and high correlation of the data points, traditional point-wise testing with multiple test adjustment suffers from very low power. It is thus desirable to seek new testing approaches that are not only more powerful but also scalable to the high dimensionality.

1.2.5 Identifying longitudinal changes of monkey's cortical measurements

In addition to the 3D ADNI TBM imaging data, in this dissertation we will also consider the analysis of functional data supported on manifolds. A motivating example is monkey brain cortical measurements. A monkey's cerebral cortex, like a human's, contains sophisticated convoluted structure with numerous concave sulci and convex gyri [47]. Such structure allows for a drastic increase of the cortical surface area in contrast to the brain skull, and enables the cortex to accommodate a greater number of neurons [71].

In this dissertation, we will consider cortical measurements, such as cortical thickness and curvature, of a group of monkeys during their early stage of development (i.e., 0-36 months). The cortical data are measured sparsely over time across different age markers. At each time marker, the cortical measurements are densely measured over the cortical surface, with 163,842 measurement points per sample. Preprocessing is conducted to align all cortical measurements of multiple monkeys over time on a common template. There are 36 monkeys in total. Each monkey has 4 - 6 measurements at different time markers. For demonstration purposes, a plot of the cortical thickness data from monkey #1 measured over four time grids is shown in Figure 1.6, in which we use colors to denote the cortical thickness.

As shown in Figure 1.7, one important characteristic of this data is that different monkeys may be measured on different time grids, thus along the longitudinal directions, the time-grids are sparse and irregular.

The goal of our analysis is to test whether asymmetric patterns exist between left and right brain, and if so, where and when the asymmetry occurs. Another point of interest is to detect whether there are increases/decreases in cortical measurements across the time of early development, and if so, where these differences occur. As the cortex surface is a 2-d

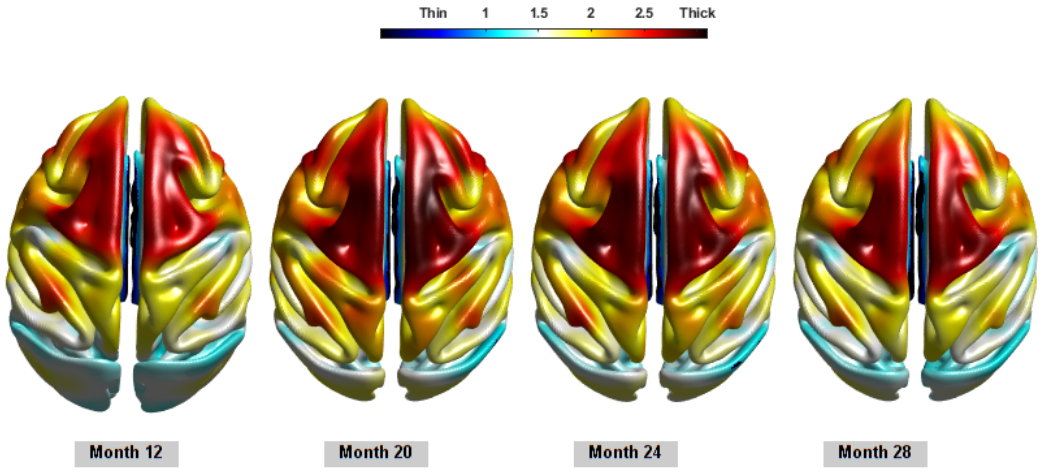

Figure 1.6: Plots of the cortical thickness of monkey #1 at its age of 12, 20, 24, and 28 months.

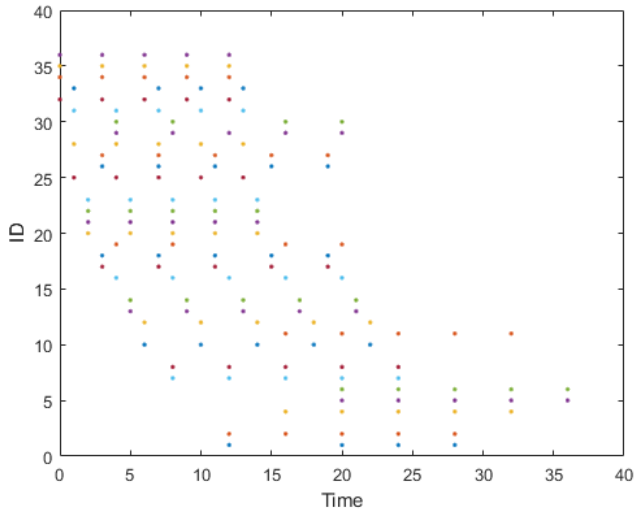

Figure 1.7: Plot of measurement time of each monkey. Different monkeys are measured on different time grids. The measurements are sparse and the time grids are irregular.

manifold embedded in 3-d space, analyzing the cortical measurements directly by ignoring the geometric structure is problematic. Therefore, we treat the cortical measurements as functional data supported on manifold and investigate appropriate representations and inference for such data. One approach is to construct an one-to-one map between the cortical

surface and a sphere by "inflating" the cortex surface to a standard sphere, on which data can be parsimoniously represented using spherical bases such as spherical wavelets. Figure 1.8 demonstrates the left (upper row) and right (bottom row) hemispheres of the cortical thickness data of Figure 1.6, plotted on standard spheres.

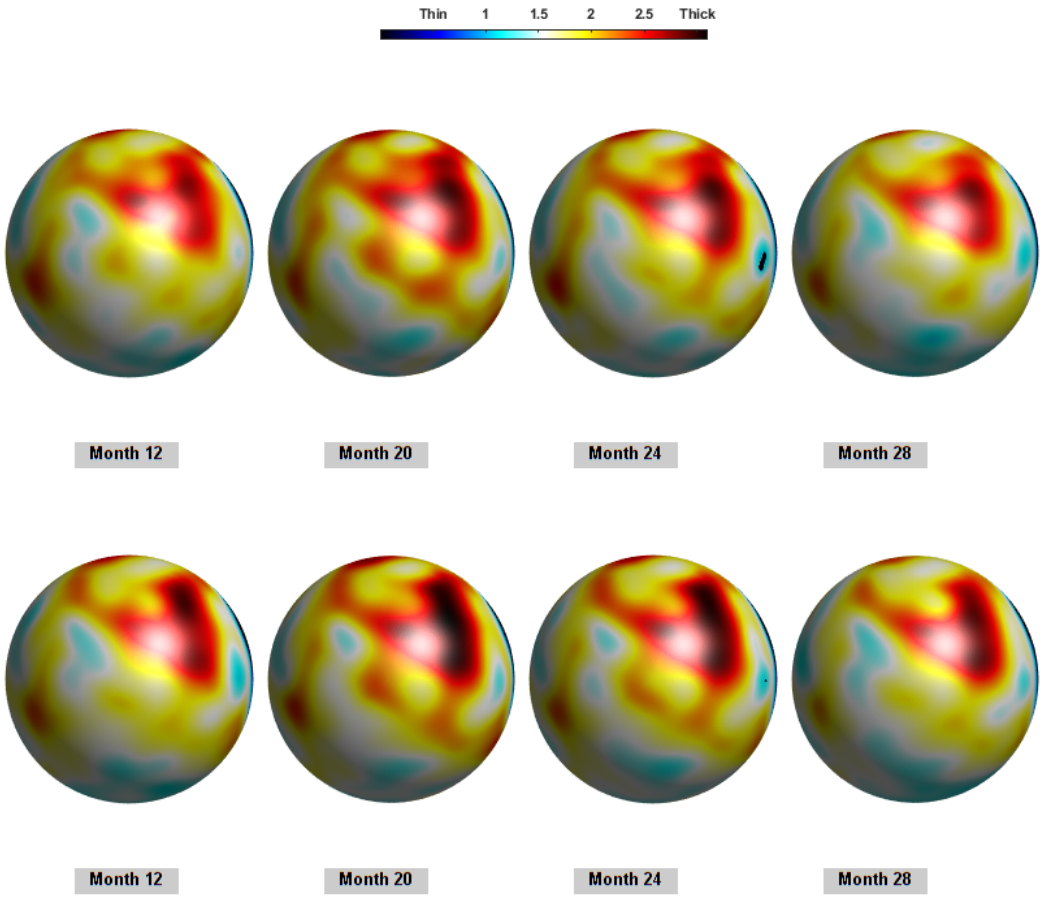

Figure 1.8: Plots of the inflated cortical surfaces of one monkey. The monkey is the same one as shown in Figure 1.6, on the spherical surface. First row shows the left brain and the second row shows the right brain. The data supported on a can be represented by parsimonious basis, such as spherical wavelets.

1.3 Background

In this section, we will review some key concepts that are relevant to the methods proposed in later chapters. These include fluctuation experiments and the generalized birth-death process in Section 1.3.1, the idea of Approximate Bayesian Computation (ABC) approach in Section 1.3.2, multiple testing with Family-wise Error Rate (FWER) and False Discovery Rate (FDR) control in Section 1.3.3, and functional principal components analysis for irregularly spaced longitudinal data in Section 1.3.4.

1.3.1 Fluctuation experiments and the generalized birth-death process

A classical problem in computational biology is to determine the spontaneous mutation rate in cultured cells. This problem, usually called fluctuation analysis, can be traced back to Salvador Luria and Max Delbrück [36] for their remarkable work that won the 1969 Nobel Prize in Physiology or Medicine. The principle of fluctuation experiments includes growing some cells (e.g., bacteria or yeast) independently to parallel cultures, plating these cultures onto a virus (lethal) agent, and counting the number of mutants that are resistant to the virus in each culture. The original purpose of fluctuation analysis is to test whether genetic mutations arise spontaneously (i.e., occur in the absence of selection), or occur as a response to selection. Following Luria and Delbrück's pioneering work (which did prove that mutations in bacteria arise spontaneously in support of Darwin's theory of natural selection), numerous research has been done to investigate the theoretical underpinning of the distribution of mutant counts in fluctuation experiments [78].

Although several mutation rate estimators have been proposed for analyzing fluctuation ex-

perimental data, including the P0 estimator [36], the median estimator [30], the modified median estimator [73], the Lea–Coulson estimator [30], the maximum likelihood estimator [28, 79], and some Bayesian estimators [2], all these methods are developed under the Lea–Coulson formulation of fluctuation experiments by assuming deterministic growth for nonmutant cells but stochastic growth (i.e., Yule process) for mutant cells. Using this formulation, the distribution of mutant counts can be shown to follow the so-called Luria Delbrück (LD) distribution:

$$p_0 = e^{-m},\ p_k = \frac{m}{k}\sum_{j=1}^{k}\phi^{j-1}(1-\frac{j\phi}{j+1})p_{k-j},\ k \geq 1$$

where

$$m = \frac{\mu}{\beta_1}(n-n_0) = \mu_\beta(n-n_0), \phi = 1-e^{\beta_1 t} = 1-\frac{n_0}{n}$$

where p_k is the probability of observing k mutants when the population starting with n_0 nonmutant cells grows to size n with cell growth rate β_1 (for both mutants and nonmutants) and mutation rate per unit time μ. Though easy to implement, such a formulation certainly cannot fit the real data which come from the more flexible (or less constrained) generalized birth-death process model. Recently, Wu and Zhu [74] developed a fast maximum likelihood estimator (MLE) for mutation rates under the Bartlett formulation of stochastic growth for both nonmutant and mutant cells. This method is based on a birth-death process model (or equivalently, a Markov branching process), and improves the computational speed substantially over the traditional MLE while allowing arbitrarily large parallel cultures and divergent culture sizes. However, due to the constraints caused by model assumptions on binary fission and non-differential growth, it still cannot meet the demand for estimating mutation parameters in complex fluctuation experimental environments.

To fill the research gap, we consider using a simulation-driven, likelihood-free estimator

(through Approximate Bayesian Computation—ABC) to estimate the mutation parameters in a generalized birth-death process. In particular, we assume that the data are from a traditional birth-death process but allow the mutation rate to be dynamic, i.e., a time-varying function. All other settings in the birth-death process keep unchanged. For the simplest case of dynamic mutation rate, we consider a piecewise constant function of time t, $p(t) = p_1\mathbf{1}_{0<t\leq\tau} + p_2\mathbf{1}_{\tau<t\leq 9}$. Note that unlike in the traditional birth-death process, for this case, the explicit likelihood of observing the nonmutant and mutant counts at a given time point cannot be derived. Thus, all existing methods for mutation rate estimation would fail. However, by comparing the observed real data with data from extensive forward simulations, we hope to inversely find the best configuration of the parameters: p_1, p_2 and τ by using the ABC estimator. We also note that, in general this proposed method would allow more complex situations such as arbitrary shape of mutation rate function $p(t)$ or differential growth of mutant/nonmutant cells, however we only limit the scope of this study to piecewise constant mutation rate function to reduce the ill-posedness caused by model nonidentifiability.

Another issue is the cost of forward simulation. For a typical fluctuation experiment in bacteria, the size of one cell culture may contain $10^8 \sim 10^{10}$ cells after 5 days growth. One simulation of such a cell culture will take about 7 seconds. Since the ABC algorithm depends on huge number of simulations, when embedding the forward simulation procedure to the ABC algorithm to obtain the posterior samples, even parallel computation on multi-core high-performance computers becomes too expensive. This motivates us to seek for an accelerated version of the ABC algorithm which could: (1) provide us a solution to likelihood-free estimation of the unknown parameters; (2) provide joint posterior distribution of all underlying parameters, which is otherwise intractable by using other analytical methods; and (3) provide speedy computation and better mixing by using a surrogate model, and thus

allow even higher dimension of the parameter space.

In forward simulation (either constant or dynamic mutation rate), we generate the total number of cells and the mutant counts by time t based on a generalized birth-death process (by treating mutant as "death"). This forward simulation procedure is summarized in Algorithm 1. Note that in this simulation model t is a known parameter, which is the checking time point and a, the exponential life span parameter, is a nuisance parameter, which is always set to be 1. If we further generalize this simulation model to incorporate a time-varying mutation rate $p(t)$. The explicit form of MLE will be hard to find. However, it is pretty easy for us to implement the simulation model, which motivated us to find a likelihood free and simulator-based estimator.

Algorithm 1 The birth-death process model:

Inputs: a, p, t
Step 1. Calculate the lifetime T of the first generation, which follows exponential distribution with rate $1/a$. Initialize culture size N_t by the number of cells with T greater than t and number of mutant cells $X_t = 0$
Step 2. Count X_t and N_t up to time t
while $T < t$ **do**
Update lifetime T by $T + T_new$, $T_new \sim Exp(1/a)$
Cells with $T < t$ will mutate with probability p if the parent is non-mutant cell, with 1 if the parent is mutant cell. Mark the mutant cells
Update $N_t = N_t +$ number of cells with $T > t$, $X_t = X_t +$ number of mutant cells
end while

1.3.2 Review of ABC algorithms

Rejection-based ABC Let Y denote a random element whose realizations are the observed data and let θ denote a parameter that determines the distribution of Y. In a typical Bayesian setup, one computes the posterior distribution $\pi(\theta|Y) \propto \pi(Y|\theta)\pi(\theta)$, where $\pi(Y|\theta)$ is the likelihood that relates Y to the parameter θ and $\pi(\theta)$ is the prior distribution for θ.

Approximate Bayesian Computation (ABC), initially proposed by Pritchard et al. [51], aims to approximate the posterior distribution $\pi(\theta|Y)$ without explicitly specifying the likelihood $\pi(Y|\theta)$. In particular, we assume that $\pi(Y|\theta)$ is unknown, but there is a simulation model, often denoted by $\pi(X|\theta)$, that produces simulated data (pseudo-data) X given θ^*. Here, θ^* is an arbitrary sample from the prior distribution $\pi(\theta)$. If X is "close to" Y, we retain θ^* as a sample of θ, otherwise, we reject θ^* and repeat the procedure with a new θ^*. This procedure will be repeated until the desired amount of "good samples" are collected. In ABC, we often use a distance measure $\rho(\cdot,\cdot)$ to determine how close X is to Y. For example, in the univariate case, by letting $\rho(X,Y) = |X-Y|$, we will retain θ^* when $|X-Y| \leq \epsilon$ for a small ϵ.

The above procedure indeed produces samples for the distribution $\pi(\theta|\{\rho(X,Y) \leq \epsilon\})$, a distribution that is identical to $\pi(\theta|Y)$ when $\epsilon = 0$ (i.e., $X = Y$). However, since {X=Y} happens with probability 0 for continuous random variables, in practice we can only require $\rho(X,Y) \leq \epsilon$ for a small discrepancy ϵ, which results in $\pi(\theta|\{\rho(X,Y) \leq \epsilon\})$. The distribution $\pi(\theta|\{\rho(X,Y) \leq \epsilon\})$ serves as an approximation of $\pi(\theta|Y)$ when ϵ is small, i.e.,

$$\pi(\theta|Y) \approx \pi(\theta|\rho(X,Y) \leq \epsilon), \text{ for a small } \epsilon.$$

When we have multiple observations $Y_i, i = 1, ...n$, we denote $\mathbf{Y} = Y_1, ..., Y_n$. We could also obtain multiple samples $\mathbf{X} = X_1, ..., X_m$ from the simulation model at one given θ^*. Then the discrepancy function ρ will be defined on a summary statistic of the samples. Let S($\mathbf{Y}$) be the sufficient statistic for θ. We have $\pi(\theta|\mathbf{Y}) = \pi(\theta|S(\mathbf{Y}))$ and $\pi(\theta|S(\mathbf{Y}))$ could be further approximated by $\pi(\theta|\rho(S(\mathbf{X}), S(\mathbf{Y})) < \epsilon)$. For example, if our data $Y_1, ..., Y_n$ is a sequence of observations from a fluctuation experiment. $Y_i = (N_{ti}, Z_{ti})$, where N_{ti} and Z_{ti} are the number of overall cells and mutant cells respectively for the ith culture up to time

t. As we have discussed in Section 1.3.1, based on the birth-death process, the estimators are proposed on the summary statistic $(\bar{N}_t, \bar{Z}_t)$. Thus, we could define $S(\mathbf{Y}) = (\bar{N}_t, \bar{Z}_t)$ and define $\rho(S(\mathbf{X}, S(\mathbf{Y})) = |S(\mathbf{X}) - S(\mathbf{Y})|$. The above procedure is summarized in Algorithm 2.

Algorithm 2 Rejection-based ABC algorithm

Inputs: $\mathbf{Y}$, ρ, Δ, ϵ ,$\pi(\theta)$,m, N, the simulator.
Initialize n = 0
while $n < N$ **do**
  Propose θ^* from $pi(\theta)$.
  Generate $\mathbf{X}$ as m samples from the simulator at θ^*
  Calculate Summary statistics S($\mathbf{X}$) from $\mathbf{X}$
  if $\rho(S(\mathbf{X}), S(\mathbf{Y})) < \epsilon$ **then**
    Save θ^* and update n= n+1;
  end if
end while

Markov Chain Monte Carlo for ABC The idea of the traditional ABC is intuitive. It relies on accepting θ^* when $\rho(S(\boldsymbol{X}), S(\boldsymbol{Y})) \leq \epsilon$. But it is inefficient for two reasons: (1) the good sufficient statistic is hard to find, and most of the times, one has to use the original data as the sufficient statistic. (2) The acceptance rate could be extremely slow as it has no memory of the previous samples. At each time it proposes a number directly from the prior. The efficiency will be especially low when the dimension of parameter is high (due to curse of dimensionality). To overcome this, several algorithms have been proposed. Here we look at one variation which incorporates the idea of Metropolis-Hasting (MH) sampler. The algorithm is summarized in Algorithm 3. More discussions of the Markov chain Monte Carlo (MCMC) variations for ABC can be found in recent literature ([69], [41], [16] and [55]).

The main idea of the MCMC-based ABC is to transfer the small discrepancy criterion to the acceptance probability controlled by the discrepancy parameter ϵ by assuming a distribution of the summary statistic. For example, we can assume that S($\mathbf{Y}$) follows a multivariate

Algorithm 3 The MCMC-ABC algorithm

Inputs: $\mathbf{Y}$, θ, ϵ, $q(\cdot \mid \cdot)$, $\pi(\theta)$, m, N, the simulator.
for $i = 1$ to N **do**
  Propose θ^* from $q(\theta^* \mid \theta)$.
  Generate m samples at θ^* and m samples at θ
  Calculate the acceptance rate following 1.3.
  Generate a random number u for $U(0, 1)$
  if $u < \alpha$ **then** Set $\theta = \theta^*$ and save θ.
  elseSave θ
  end if
end for

Gaussian distribution with diagonal matrix and write:

$$\pi_\epsilon(S(\mathbf{Y})|S(\mathbf{X})) = (2\pi\epsilon)^{J/2} exp\{-\frac{1}{2\epsilon^2}(S(\mathbf{Y}) - S(\mathbf{X}))^T(S(\mathbf{Y}) - S(\mathbf{X}))\}, \tag{1.1}$$

where J is the dimension of the sufficient statistic. We then could approximate $\pi(S(\mathbf{Y})|\theta)$ by $\pi_\epsilon(S(\mathbf{Y})|\theta)$ by Monte Carlo integration, which is:

$$\pi_\epsilon(S(\mathbf{Y})|\theta) = \int \pi_\epsilon(S(\mathbf{Y})|S(\mathbf{X}))\pi(S(\mathbf{X})|\theta)dS(\mathbf{X}) \approx \frac{1}{H}\sum_{g=1}^{H} \pi_\epsilon(S(\mathbf{Y})|S(\mathbf{X}^{(g)})). \tag{1.2}$$

Here, we have H samples from the simulation model and denote each as $\mathbf{X}^{(g)}, g = 1, ..., H$. Note that $\pi(S(\mathbf{X})|\theta)$ could be generated from the simulator directly. Suppose the proposal distribution is $q(\theta^*|\theta)$. We will then accept the new proposed θ^* with probability

$$\alpha(\theta^*|\theta) = min\{1, \frac{\pi(\theta^*)\pi_\epsilon(S(\mathbf{Y})|\theta^*)q(\theta|\theta^*)}{\pi(\theta)\pi_\epsilon(S(\mathbf{Y})|\theta)q(\theta^*|\theta)}\}. \tag{1.3}$$

The above MCMC procedure has improved mixing of the posterior samples, much better than the traditional rejection-based one, because it is not blindly sampling from the prior. However, it needs to call the simulator H times in each step to evaluate the acceptance ratio. Here, H needs to be large enough to guarantee a good approximation, e.g., $H = 1000$ is

reasonable for a Gamma simulator. It may bring computational burden when the simulator runs slow. It cannot take into account the possibility of making a wrong decision. Thus, to obtain a good sample, we may need more iterations. Furthermore, the samples are only used to make one decision and then thrown away. In fact, we hope to remember the samples somehow instead of calling the simulator repeatedly. In Chapter 2, we adopt a Gaussian process surrogate (GPS) for the simulator following the idea of Mee [1], (discussed in Section 2.1.2), which substantially reduces the number of simulation calls.

1.3.3 Multiple testing

Simultaneous inference was introduced as a statistical problem as early as the mid-twentieth century. It has been reactivated recently because the advances in technology have provided us more data sets containing large number of variables with more complex structures, for example the data shown in Figure 1.4, 1.5, and 1.6. They are in the forms of curves, images or even manifold. For this reason, testing procedure for such high dimensional/complex structured data is of interest. In this section, we will review some of the major contributions to multiple hypothesis testing.

Hypothesis Testing Before reviewing multiple testing methods, we recall some basics of hypothesis testing. We typically set our the hypotheses as:

$$H_0 : \theta \in \Theta_0 \; vs. \; H_1 : \theta \in \Theta_1.$$

We often define a test function $\phi(.)$ to map the data $\mathbf{X}$ onto the range [0,1] in order to make decision on whether to reject the null hypothesis H_0. There are two types of error that we may make: Type I error occurs when we decide to reject the null when it is in fact true;

Type II error occurs when we fail to reject the null when the alternative is true. We control these two errors by

- controlling the Type I eror probability by a threshold α, so that

$$\sup_{\theta \in \Theta_0} E[\phi(\mathbf{X})] \leq \alpha \; or \; \lim_{n \to \infty} \sup E_\theta[\phi(\mathbf{X})] \leq \alpha;$$

- and maximizing the power on Θ_1:

$$\beta(\theta) = E_\theta[\phi(\mathbf{X})] \; for \; \theta \in \Theta_1.$$

Note that it is not generally possible to maximize power function $\beta(\theta)$ uniformly over Θ_1, but that is ideal and sometimes achieved (e.g, using Karlin-Rubin Theorem)

The testing methods are controlling errors by p-value. Here, we provide a formal review of the p-value definition. For a hypothesis of the form:

$$H_0 : \theta \leq \theta_0 \; vs. \; H_1 : \theta \geq \theta_0$$

and testing function $\phi(\mathbf{X}) = \mathbf{1}_{(T(\mathbf{X}) > C_\alpha)}$, where C_α is the critical value that is determined so that

$$\sup_{\theta \in \Theta_0} = \sup_{\theta \in \Theta_0} Pr(T(\mathbf{X}) > C_\alpha) \leq \alpha,$$

which implies that $Pr(T(\mathbf{X}) > C_\alpha \mid H_0) \leq \alpha$. We define p-value by

$$p = Pr(T(\mathbf{X}) > t(\mathbf{x}) \mid H_0),$$

where $t(\mathbf{x})$ is the observed value of $T(\mathbf{X})$. We see that $t(\mathbf{x}) > C_\alpha$ iff $p < \alpha$. So if one knows the p-value, one can make decision simply by comparing it with α.

Multiple testing For the testing problem motivated by the examples in Section 1.2.3, 1.2.4 and 1.2.5, there are thousands or even more than a million pointwise tests. Suppose in total we have m tests:

$$H_{0,1} \; vs. \; H_{a,1}, ..., H_{0,m} \; vs. H_{a,m}.$$

Here, we assume they are independent of each other. If we still control each single hypothesis's type I error at α in the way described in a single hypothesis, the overall type I error will approach 1 as m increases. Thus, we need testing methods that can control the overall level of Type I error. In order to realize that, the decision rule needs to be adjusted. Many multiple testing methods are proposed by adjusting the p-value, and call the value obtained the adjusted p-value. With adjusted p-value, one could simply make decision by comparing the adjusted p-value to the level α.

There are mainly two types of controls for multiple testing methods, one is called Family-wise Error Rate control and the other is known as False Discovery Rate control.

1. Family-wise Error Rate Control

This approach is to control the overall family-wise error rate to be less than or equal to α. Denote $\cap_i H_i$ the event that all $\{H_{0,i}, i = 1, ..., m\}$ hold, and denote $\cap_{i \in I} H_{0,i}$ the event that a subset $\{H_{0,i}, i \in I\}$ hold, $I \subseteq \{1, ..., m\}$. There are two types of family-wise error rate that can be controlled:

- FWE under complete null hypothesis (FWEC),

$$FWEC = Pr(Reject \; at \; least \; one \; H_{0,i} \mid \cap_i H_{0,i}).$$

- FWE under partial null hypothesis (FWEP),

$$FWEP = Pr(Reject\ at\ least\ one\ H_{0,i} \mid \cap_{i\in I} H_{0,i}).$$

We say a multiple testing approach achieves weak control of the family-wise error if $FWEC \leq \alpha$; and achieves strong control of the family-wise error if $FWEP \leq \alpha$ for all $I \subseteq \{1, .., m\}$. The commonly used multiple testing corrections that control FWE are introduced below, in the way of the expression of adjusted p-value.

1. Boferroni Correction:

   Denote p_i the p-value for the ith hypothesis $H_{0,i} vs. H_{a,i}$. The Boferroni correction will reject $H_{0,i}$ the null when $p_i < \alpha/m$. Equivalently, one can define adjusted p-value for the ith hypothesis by

$$\tilde{p}_i = \min(mp_i, 1) \tag{1.4}$$

   and reject $H_{0,i}$ if $\tilde{p}_i < \alpha$. By assuming that the tests are independent, one can show that Bonferroni correction could control FWE in the strong sense. The biggest issue of it is that it is too conservative. Suppose we have 1000 tests and the α level is 0.05, one test should have an unadjusted p-value as small as 5×10^{-5} to be rejected. When there are more tests, it is even harder to reject a test. So that the power of this method is very low. Moreover, the assumption of independence is not the case for functional data.

2. Šidák Method

   The Šidák will reject H_i when $p_i \leq 1-(1-\alpha)^{1/m}$. The corresponding adjusted p-value for the ith hypothesis is

$$\tilde{p}_i = 1 - (1 - p_i)^m. \tag{1.5}$$

This correction provides FWEC at exact α level if all tests are independent and p-values are distributed as $Unif(0,1)$. If p-values are not independent, it will control $FWE \leq \alpha$

3. Holm and Hochberg

   The Bonferroni and Šidák methods are called single-step methods because they perform adjustments for all tests simultaneously, regardless the order of all the unadjusted p-values $p_1, .., p_m$. There is another family of step-wise method that allow different adjustments for different hypotheses, depending on how the original p-values are ordered. One example is the method proposed by Holm and Hochberg, coined as "Holm's method".

Let $\{p_i, i = 1, ..., m\}$ be the p-values for the m hypotheses. We first order these p-values in non-decreasing order:

$$p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)},$$

where $p_{(i)}$ denotes the ith smallest p-value. We denote the corresponding null hypotheses:

$$H_{0,(1)}, H_{0,(2)}, ..., H_{0,(m)}.$$

It improves Bonferroni correction by adjusting p-values step by step from the smallest one follwing the steps:

Step 1: We stat with $p_{(1)}$. If $p_{(1)} < \alpha/m$, reject $H_{(0,1)}$ and continue. Otherwise, stop and accept all $H_{(0,i)}, i = 1, .., m$.

Step 2: For the rest of $m - 1$ hypotheses, we apply Bonfferoni correction, i.e, If $p_{(2)} <$

$\alpha/(m-1)$, reject $H_{(0,2)}$ and continue. Otherwise, stop and accept all $H_{(0,i)}, i = 2, .., m$.

$\vdots$

Step j: If $p_{(j)} < \alpha/(m-j+1)$, reject $H_{(0,j)}$ and continue. Otherwise, stop and accept all $H_{(0,i)}, i = j, .., m$.

$\vdots$

Step m: If $p_{(m)} < \alpha$, reject $H_{(0,1)}$. Otherwise, stop.

To summarize, we reject $H_{(0,j)}$ iff $p_{(j)} < \alpha/(m-j+1)$ for all $i = 1, ..., j$. This is equivalent to adjust the p-values to:

$$\begin{aligned}
\tilde{p}_{(1)} &= \min(mp_{(1)}, 1) \\
\tilde{p}_{(2)} &= \max\{\tilde{p}_{(1)}, \min((m-1)p_{(2)}, 1)\} \\
&\vdots \\
\tilde{p}_{(j)} &= \max\{\tilde{p}_{(j-1)}, \min((m-j+1)p_{(j)}, 1)\} \\
&\vdots \\
\tilde{p}_{(m)} &= \max\{\tilde{p}_{(m-1)}, p_{(m)}\}
\end{aligned}$$

Here, maximum is used in each step to ensure the monotonicity of the adjusted p-values as well as ensure the order of rejection. In summary, the adjusted p-value corresponding to $H_{0,(j)}$ is

$$\tilde{p}_{(j)} = \max_{i=1,...,j} \min((m-i+1)p_{(i)}, 1) \tag{1.6}$$

The Holm's method controls FWER in the strong sense. It is less conservative than Bon-

ferroni. All tests rejected after Bonferroni correction will be rejected by Holm's method but not vice versa.

2. False Discovery Rate control

FWER control is one criterion of controlling the overall level of error. It generalizes the definition of type I error under the multiple testing's setup. However, as the control is strong, it is always too conservative to discover interesting effect when the number of tests is large. Thus, another criterion is proposed to deal with this, which is called False Discovery Rate (FDR) control. In Table 1.1, we list the configurations of performing m hypotheses simultaneously. Note that only m and R are known to us.

	Accept$H_{0,i}$	Reject $H_{0,i}$	Total
true $H_{0,i}$	U (correct decision)	V (Type I error; FP)	m_0
true $H_{a,i}$	T (Type II error;FN)	S (correct decision)	m_1
total	$m - R$	R	m

Table 1.1: Confusion table for multiple testing with notation sets: 1. m: total number of tests; m_0: total number of true null; m_1: total number of true alternatives. 2. U: number of accepted true null hypotheses; V: number of rejected true null hypotheses. 3. T: number of true alternatives that are failed to reject; S: number of true alternatives that are correctly rejected. 4.R: number of tests that were rejected.

Under the complete null hypothesis, we have $m_0 \equiv m$ whereas under the partial null hypothesis, we have $m_0 < m$. Under this configuration, the FWER and FDR are defined as follows:

1. The family-wise error rate is the probability of rejecting at least on $H_{0,i}$ given that $H_{0,i}$ holds. It is defined by

$$FWER = Pr(V \geq 1)$$

2. The false discovery rate is the expected proportion of false rejection. Here, we define

the proportion of false rejection by

$$Q = \begin{cases} V/R & if\ R > 0 \\ 0 & if\ R = 0 \end{cases}$$

Thus, FDR = E[Q]. We could show that FDR is not larger than FWER. Thus, the method controls FWER also controls FDR. Benjamini and Hochberg [3] presented the first procedure for controlling FDR (BH algorithm). It still remains the most common procedure to date.

1. BH algorithm:

   The procedure is described in details as follows:

   Step 1 Get ordered p-values $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$.

   Step 2 Compare each p-value $p_{(i)}$ with its corresponding threshold $i\alpha/m$.

   Step 3 Let $k = \max(k : p_{(k)} \leq \frac{k\alpha}{m})$, reject all $H_{(0,i)}, i = 1, 2, ..., k$

   To summarize, given α, we'll find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$, and reject all $H_{(0,i)}$ with $i \leq k$.

   For independent test statistics and for any configuration of false null hypotheses, the BH algorithm controls the FDR at α. The corresponding adjusted p-values are calculated as:

$$\tilde{p}_{(i)} = \min_{k=1,...,m} \{\min(\frac{m}{k} p_{(k)}, 1)\} \tag{1.7}$$

Based on the background above, we propose a testing procedure which is conducted in basis space with p-value guided compression in Chapter 3. It could be incorporated with both FWER and FDR correction methods. With Westfall-Young randomization adjustment procedure, it does not require assumption of independence of tests and distribution of data.

It is scalable, so that we could easily parallelize the algorithm to deal with high dimensional data and the property keeps stable even when the number of tests approaches to inifinity. We show that: (1) the Westfall-Young randomization test in basis space achieves strong control of FMER and asymptotic optimality, and (2) by applying appropriate basis representation and compression, the power could be further improved than its point-wise counterpart. Details will be discussed in Chapter 3.

1.3.4 Functional principal components analysis for irregularly spaced longitudinal data

Recall the example introduced in Section 1.2.5, we have sparsely and irregularly collected measurements on a manifold surface for each monkey. In order to apply the basis-space testing procedure in Chapter 3 to test the asymmetry and trend across time, we need to estimate the whole trajectory of each spherical wavelet component. For that purpose, we consider estimating the trajectory by a set of functional principal component scores obtained from data through functional principal component analysis.

Functional principal component analysis is a commonly used tool to reduce the dimension of data in functional form. In addition, it could also characterize the parts of the curves with most variation around an overall mean trend. There is plenty of literature on this topic and the summary could be found in Ramsay and Silverman [53]. Specifically, for irregular repeated measurements, Staniswalis and Lee [61] has proposed a kernel-based method. However, when the functional principal component scores are defined through *Karhunen-Loève* expansion, the usual integration method cannot approximate properly. Thus, they require dense measurements. There is also considerable literature dealing with sparse data. One widely investigated approach is through mixed effect models, such as the works of James

et al. [24] and James and Sugar [25]. The drawback of this approach is that it is too complex to derive nice asymptotic properties.

In order to address both sparsity and irregularity of our data properly, we use a technique called Principal Components Analysis Conditional Expectation (PACE), proposed by Yao et al. [77]. This is a method designed to handle sparse and irregular longitudinal data. It has nice asymptotic properties as long as the pooled time markers is sufficiently dense. However, it allows down to one or two measurements for each subject or at one time marker. They also provide guide to choose the number of eigenfunctions, such as one-curve-leave-out cross validation and a faster way of using AIC criterion. The main idea of this approach is to borrow information from other time markers to estimate the eigenfunctions. Next, we will briefly review the model and the estimation procedure of this method.

Let the observed data be Y_{ij}, $i = 1, \ldots, n; j = 1, \ldots N_i$, where N_i's are i.i.d random variable indicating the sparse and irregular pattern of the data. It could be modeled as $Y_{ij} = X(t_{ij}) + \epsilon_{ij}, t_{ij} \in \mathcal{T}$, where $X(t_{ij})$ is a random function with mean function $\mu(t)$ and covariance function $G(s,t) = cov(X(s), X(t))$. $\mathcal{T}$ is the domain of time marker t, which is a closed interval in R. We could also assume the data is a function over a space, such as an image, in which case $\mathcal{T}$ is a subspace of R^2. ϵ_{ij} stands for the measurement error of the jth measurement for the ith subject. Assume that there exists a set of orthogonal eigenbasis in $\mathcal{T}$ and the covariance function could be expanded as $\phi_k(t), k = 1, \ldots, \infty$. That is, we could find a non-increasing sequence of eigenvalues λ_k, such that, $G(s,t) = \sum_k \lambda_k \phi_k(s)\phi(t)$. We could then represent the random curves $X(t_{ij})$ in terms of the eigenfunctions as $X(t_{ij}) = \mu(t_{ij}) + \sum_k c_{ik}\phi_k(t_{ij})$. The mean of the random coefficient c_{ik} is 0 and the variance of it is λ_k. Thus, the final model for the observed data is

$$Y_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} \phi_k(t_{ij}) c_{ik} + \epsilon_{ij}.$$

In order to predict the whole trajectory at all time markers for one subject, we need to estimate the unknown mean function $\mu(t)$, the eigenfunctions $\phi_k(t), k = 1, ..., K$ and coefficients $c_{ik}, i = 1, \ldots, n; k = 1, \ldots, K$. Note that instead of dealing with the theoretical infinite dimension of the eigenfunctions, we will use only limited number of eigenfunctions to predict the trajectory. The number of basis used could be determined by cross-validation or AIC criterion. To keep things simple, we will not go into details of the estimation procedure. Interested readers could find the details in Yao et al. [77]. In short, the mean function will be first estimated by a local linear smoother based on the entire pooled data set. The sparseness will bring trouble in traditional integration method for estimating component scores. Hence, they consider to estimate the scores using conditional expectation under Gaussian assumptions. The quantities used to estimate the conditional expectation could be derived from the pooled data as well. Thus, we could borrow the information from subjects at other time markers through estimation of mean and covariance function, making the estimation under sparse and irregular condition possible.

Later, in Chapter 4, we will use this technique to achieve trajectory of each spherical wavelet component. It allows us to make best use of the data and makes it possible to align the sparsely and irregularly collected data for later testing procedures.

1.4 Outline of the Dissertation

We introduce scalable parameter estimation for complex, high-dimensional data in Chapter 2. In Chapter 3, a randomization-based multiple testing approach will be proposed for high-dimensional data in functional form. The analysis of functional data supported on manifold will be introduced in Chapter 4. A summary and discussion of these statistical methods is provided in Chapter 5.

Chapter 2

Scalable Parameter Estimation for Complex, High-Dimensional Data

We consider two types of estimation problems in this chapter — the mutation rate estimation problem in fluctuation analysis and the parameter estimation problem for functional data arising from complex systems. We solve both problems by introducing an improved Approximate Bayesian Computation (ABC) estimator called Gaussian Process Surrogate ABC (GPS-ABC). This new estimator incorporates Gaussian Process Regression model to replace the real simulator, which dramatically reduces time to obtain a large number of MCMC samples and achieves better mixing than traditional ABC. We will first describe the proposed algorithm under the mutation rate estimation setup in Section 2.1.2, and then describe two algorithms under the GPS-ABC framework — the ABC estimator of mutation rate based on "birth-death" process (ABC-BD) and the ABC estimator of mutation rate function based on "generalized birth-death" process (ABC-GBD).

In Section 2.2, we further extend the GPS-ABC to functional outputs by introducing wavelet decomposition and compression. We call the proposed estimation method wavelet-based ABC (wABC). The performance of wABC is shown in the foliage-echo application.

2.1 Using the Approximate Bayesian Computation approach to estimate mutation rate

2.1.1 Introduction

Estimating mutation rate is a classical problem in fluctuation analysis. Traditional methods assume a distribution for the mutant counts and rely on estimators that are based on the likelihood function. Though easy to implement, these estimators can hardly be generalized to incorporate non-constant mutation rate. Another interesting approach is to assume stochastic growth for both nonmutant and mutant cells based on a birth-death process model. This method improves the speed of estimating MLE over the traditional estimator and provides a way to simulate fluctuation experiment. However, it is still likelihood-based method and thus not flexible enough to meet the demand in complex fluctuation experimental experiments. Approximate Bayesian computation, due to its flexibility in estimation without relying on an explicitly specified likelihood function, provides a promising alternative for estimating parameters in complex birth-death processes. In this section, we introduce a variation of ABC algorithm called GPS-ABC under the birth-death process model setup in Section 2.1.2. The proposed new estimator of mutation rate is called ABC-BD. We further generalize the simulation model to incorporate time-varying mutation rate and apply it to estimate parameters in Section 2.1.3. We conduct two simulation studies for small scale case (population size at level 10^5) in Section 2.1.4, and apply the proposed estimators on a published large scale experiment dataset in Section 2.1.4.

2.1.2 GPS-ABC estimator for birth-death process model

Simple ABC algorithms, such as ABC and MCMC-based ABC (described in Section1.3.2), suffer from slow-mixing and are not scalable to high dimensionality, because they either search blindly in the whole parameter space or needs to call the simulation models many times to calculate one acceptance rate. In order to further improve the efficiency and mixing of the ABC algorithm, Meeds and Welling [41] proposed Gaussian Process Surrogate ABC (GPS-ABC). The main idea is to treat the observed data or its summary statistics as response surfaces of the parameters, and use the Gaussian Process to model the relationship between them. The GP surrogate model can be used to replace the real simulator in MCMC-based ABC, which will dramatically reduce the computation time when the simulator is relatively expensive.

Gaussian process surrogate (GPS) Due to its flexibility, Gaussian process may serve as a wonderful surrogate to expensive simulation models. Instead of directly generating samples from the simulator of a complex system, we can obtain samples from a trained GP model, which is more computationally efficient. In the MCMC-based ABC framework, one may need to call the simulator more than 1000 times to calculate one acceptance rate, which is time-consuming even when each simulation only takes as little as 10 seconds.Therefore, we introduce GP surrogate to ABC in order to ABC improve the computation efficiency. We call the ABC model with GP surrogate GPS-ABC. We now review the GPS-ABC of Meeds and Welling [41] under the birth-death process setup. Suppose we have J parallel cultures in the experiment. Denote the number of overall cells and mutant cells at time t by $\mathbf{N}_t = \{(N_{t_1}, ..., N_{t_J})\}$ and $\mathbf{Z}_t = \{(Z_{t_1}, ...Z_{t_J})\}$ respectively. Let the initial number of cells be a. Based on N_t, Z_t and a, we hope to estimate the mutation rate p. Algorithm 1 in Section 1.3.1 summarizes the simulator for the birth-death process. It takes the following inputs: the checking time t, the initial number of cells a and the mutation rate p. It returns the

following outputs: the total number of cells and the number of mutant cells in culture k at time t, denoted by N'_{t_k} and Z'_{t_k} respectively. We repeat this simulation J times for J parallel cultures.

In an MCMC-based ABC framework, the synthetic likelihood function takes the form Equation 1.2 in Section 1.3.2. We assume that the components of sufficient statistics S($\mathbf{Y}$) of data are mutually independent and the acceptance ratio is calculated under this assumption. However, we notice that our data — $\mathbf{N}_t$ and $\mathbf{Z}_t$ are correlated. Using both of them directly will violate the independence assumption. Thus, we need to find define a new statistic, so that the relationship between the feature and the parameter p is well enough for a GP model to perform as a surrogate model. We propose several potential features based on the observed mutant proportion and the MOM estimator. To better observe the response surface of the features with respect to the parameters, we generate features on a grid of mutation rate using the simulation model and plot the mean of the features vs. mutation rate (or log of mutation rate) in Figure 2.1. From Figure 2.1, we can see that the feature $\sqrt{Z_t/N_t}$ and the parameter $log(p)$ are suitable choices due to the following reasons: (1) the curve looks smooth, in which case the GP surrogate model gives more precise out-of-sample predictions; (2) the feature is sensitive to the changes in $log(p)$ (i.e., the curve is never flat in any region of $log(p)$). Thus, we choose $\sqrt{Z_t/N_t}$ as the feature and $log(p)$ as the parameter in the ABC framework. The estimation of p can be easily transformed fro mthe estimation of $log(p)$.

With the chosen feature for each culture, we denote the sequence of features for all cultures by $\mathbf{d}_y = \{d_{y_1}, ...d_{y_J}\}$. In order to remove the random fluctuation, we further define the summary statistic by S($\mathbf{Y}$) $= \bar{d}_y = \frac{1}{J}\sum_{j=1}^{J} d_{yj}$. We assume that

$$\bar{d}_y = \bar{d}_x + e,\ e \sim N(0, \epsilon^2). \tag{2.1}$$

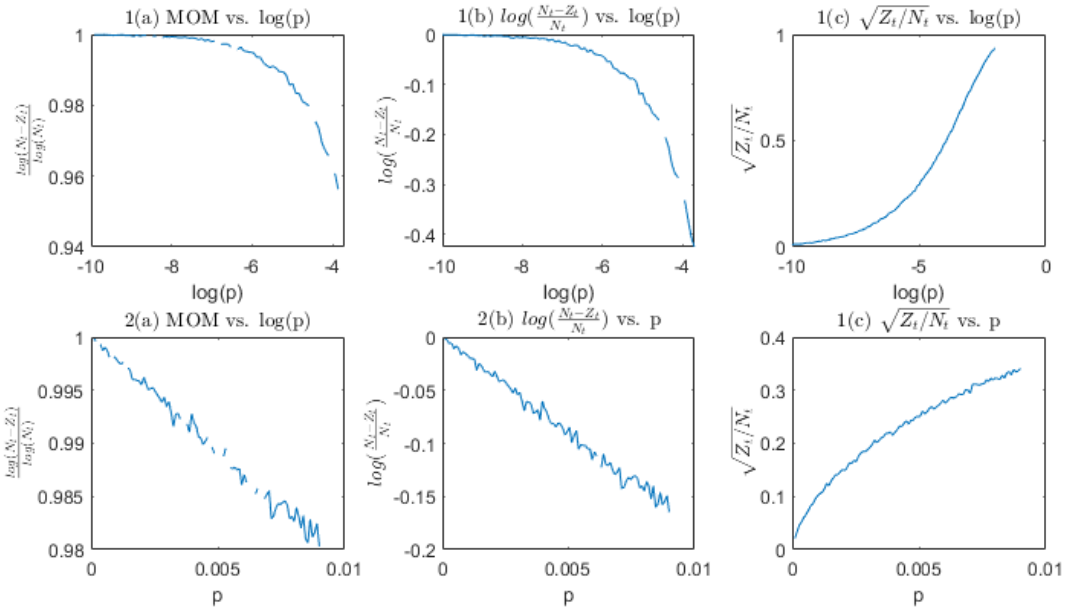

Figure 2.1: A comparison of using different features in the Gaussian process surrogate. In the first row p is in the log scale while in the second row p is in the original scale. (a) Use MOM estimator as the feature, (b) Use the logarithm of non-mutant cells proportion as the feature and (c) Use the square root of the mutant cell proportion as the feature.

Here $\bar{d}_x$ is the averaged feature value obtained from the simulated data

$$(\mathbf{N}'_t, \mathbf{Z}'_t) = \{(N'_{t_1}, Z'_{t_1}), ..., (N'_{t_M}, Z'_{t_M})\},$$

and Model 2.1 is equivalent to assuming that $\pi_\epsilon(\mathrm{S}(\mathbf{Y})|\mathrm{S}(\mathbf{Y})) = \pi_\epsilon(\bar{d}_y|\bar{d}_x)$ corresponding to a $N(\bar{d}_x, \epsilon^2)$ distribution. The simulator denoted by $\pi(\bar{d}_x|\theta)$ now generates summary statistics using input parameter $\theta = log(p)$. For large scale simulator, i.e, when checking time t is large, it takes more than 10 seconds to generate one summary statistics from 10 cultures. In order to reduce the time cost of repeatedly calling the simulator, we consider to further replace the simulator by a cheaper Gaussian Process (GP) regression model:

$$\bar{d}_x = f(\theta) + r, \;\; f(\theta) \sim GP(0, k(\theta, \theta^*)), \;\; r \sim N(0, \sigma^2) \tag{2.2}$$

where $f(\theta)$ is a GP with zero mean and covariance kernel $k(\theta, \theta^*)$. The most commonly used kernel is the squared exponential kernel $k(\theta, \theta^*) = \phi^2 exp\{-||\theta - \theta^*||^2/(2\tau^2)\}$. The GP regression model includes three hyperparameters: scale ϕ, length-scale τ and nugget σ^2. These hyperparameters could be estimated based on the training samples through MLE estimation. Details can be found in Binois et al. [4]. This GP regression model 2.2 is called a Gaussian Process Surrogate . We use it instead of the true simulator to generate $\bar{d}_x$. In Figure 2.2, we show the performance of the GP regression model compared to the true birth-death process simulator. Figure 2.2 is obtained by first taking 100 equally spaced grid points on the domain of θ - [-10,-2], generating 10 samples at each grid point, averaging the 10 samples, and using the 100 averaged values as the training samples for a GP regression. Hyperparameters of the GP regression were estimated using MLE. Using Equations 2.3 and 2.4, we also obtained the 95% predictive interval for $\bar{d}_x$. Figure 2.2 shows that, the 95% predictive interval is precise and covers the sampled data from the real simulation model. This verifies that the GPS may serve as a good surrogate for the real simulator in the ABC algorithm.

The GPS-ABC Algorithm The GPS-ABC involes calculating $\pi_\epsilon(\bar{d}_y|\theta)$ and $\pi_\epsilon(\bar{d}_y|\theta^*)$ using the GPS following a three-step procedure.

1. Produce a grid of values $\Theta = \{\theta_1, ..., \theta_{S_0}\}$ on the domain of θ. Here, $\theta = log(p)$, is on the domain from -10 to -2. At each grid point, generate $\mathbf{X} = \{d_{x_1}, ..., d_{x_M}\}$ using the simulator and denote their average by $\bar{d}_x$. This results in a list of "input-output" pairs $\{(\theta_i, \bar{d}_{x,i}), i = 1, ..., S_0\}$, which are initial samples used to train the GPS model.

2. For a pair of proposed values of θ, (θ^*, θ), we can calculate the GP predictive distribution on them as $N(\mu_{(\theta^*,\theta)|\Theta}, \Sigma_{(\theta^*,\theta)|\Theta})$, where

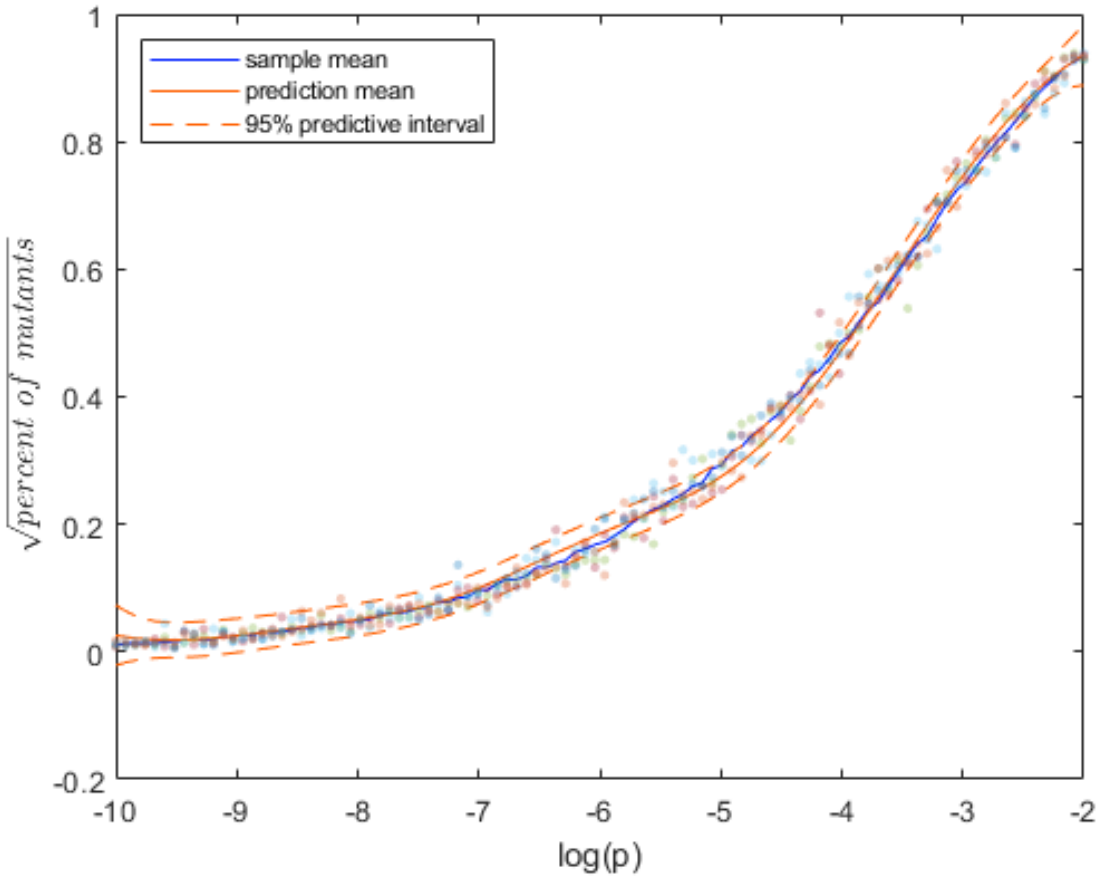

Figure 2.2: A demonstration of the GP surrogate model for the birth-death process simulation model. Here, we show the result on 100 equally spaced grid points of $\theta = log(p)$ on the domain [-10,-2]. The dots are the samples directly drawn from the simulation model from 10 times of simulations. The blue line shows the mean of the samples. The red solid line shows the GP predictive mean and the dashed line indicates the 95% predictive interval.

$$\mu_{(\theta^*,\theta)|\Theta} = \begin{pmatrix} \mathbf{k}_{\theta^*,\Theta} \\ \mathbf{k}_{\theta,\Theta} \end{pmatrix} \left(\mathbf{K}_{\Theta,\Theta} + \sigma^2\mathbf{I}\right)^{-1} \bar{\mathbf{d}}_x \ , \tag{2.3}$$

$$\Sigma_{(\theta^*,\theta)|\Theta} = \begin{pmatrix} k_{\theta^*,\theta^*} & k_{\theta^*,\theta} \\ k_{\theta,\theta^*} & k_{\theta,\theta} \end{pmatrix} - \begin{pmatrix} \mathbf{k}_{\theta^*,\Theta} \\ \mathbf{k}_{\theta,\Theta} \end{pmatrix} \left(\mathbf{K}_{\Theta,\Theta} + \sigma^2\mathbf{I}\right)^{-1} \begin{pmatrix} \mathbf{k}_{\theta^*,\Theta} \\ \mathbf{k}_{\theta,\Theta} \end{pmatrix}^T . \tag{2.4}$$

Here, $\bar{\mathbf{d}}_x = (\bar{d}_{x,1}, ..., \bar{d}_{x,S_0})^T$ is a S_0-by-1 vector of training points. $\mathbf{k}_{\theta,\Theta}$ is a 1-by-S_0 vector consisting of kernel evaluations at θ^* and components in Θ, $\mathbf{K}_{\Theta,\Theta}$ is a S_0-by-S_0 matrix constituted by kernel evaluations at every two components of Θ, and $k_{\theta,\theta^*} = k(\theta^*, \theta)$.

3. Following the MCMC-ABC framework, the likelihoods $\pi_\epsilon(\bar{d}_y|\theta)$ and $\pi_\epsilon(\bar{d}_y|\theta^*)$ can be approximated by $N(\bar{d}_y|\mu_\theta^{(l)}, \sigma^2 + \epsilon^2)$ and $N(\bar{d}_y|\mu_{\theta^*}^{(l)}, \sigma^2 + \epsilon^2)$ respectively, where $(\mu_{\theta^*}^{(l)}, \mu_\theta^{(l)})$ is a sample indexed by l from $N(\mu_{(\theta^*,\theta)|\Theta}, \Sigma_{(\theta^*,\theta)|\Theta})$. The acceptance probability of MCMC based on s samples can be calculated by:

$$\alpha^{(l)}(\theta^*|\theta) = \min\left\{1, \frac{\pi(\theta^*)N(\bar{d}_y|\mu_{\theta^*}^{(l)}, \sigma^2 + \epsilon^2)q(\theta|\theta^*)}{\pi(\theta)N(\bar{d}_y|\mu_\theta^{(l)}, \sigma^2 + \epsilon^2)\ q(\theta^*|\theta)}\right\}. \tag{2.5}$$

Here, we index the acceptance probability α by l as it is based on a random sample from the GP predictive distribution, not the true deterministic value $f(\theta^*)$ and $f(\theta)$. In other words, α is a random variable and $\alpha^{(l)}$ is one realization of it. This randomness introduces uncertainty in decision making. Thus, it will potentially influence the mixing of MCMC. Since the uncertainty of α is called by replacing the true simulator by a surrogate, we can control the uncertainty through refining the surrogate by adding training samples to the GPS while controlling the probability of making a wrong decision. Adding this step at each MCMC iteration will guarantee that every decision is made with a given level of confidence, which brings a better mixing with only limited number of callings from the real simulation model. Specifically in each iteration, we produce L samples of $(\mu_{\theta^*}^{(l)}, \mu_\theta^{(l)})(l = 1, ..., L)$ and

obtain L acceptance probabilities $(\alpha^{(1)}(\theta^*|\theta), ..., \alpha^{(2)}(\theta^*|\theta))$. In order to make decision, we need to summarize information of all α's into a summary statistic denoted by ζ. Then when making decision, we generate a random number $u \sim \text{Unif}(0, 1)$, we accept θ^* if $u < \zeta$ and otherwise reject it.

Now we can calculate the probability of making an incorrect decision. If $u < \zeta$, we accept θ^*. This decision is incorrect if actually $\{u > \alpha\}$. The probability is $\mathbb{1}_{\{u<\zeta\}} Pr(\{u > \alpha\})$. Similarly, we reject θ^* if $u \geq \zeta$ and this decision is incorrect if actually $\{u \leq \alpha\}$. The probability of this situation is $\mathbb{1}_{\{u\geq\zeta\}} Pr(\{u \leq \alpha\})$. Therefore, given a value of u, the overall probability of making error is

$$W_u(\alpha) = \mathbb{1}_{\{u<\zeta\}} Pr(\{u > \alpha\}) + \mathbb{1}_{\{u\geq\zeta\}} Pr(\{u \leq \alpha\}), \quad u \sim \text{Unif}(0, 1).$$

We can further integrate u out and obtain the marginal probability of making a wrong decision, i.e.,

$$W(\alpha) = \int_0^1 W_u(\alpha) du.$$

The above error probability $W(\alpha)$ is minimized when $\zeta = median(\alpha)$; a detailed argument can be found in the Section 3.1 of Meeds and Welling [41] and the reference therein.

To implement the control of uncertainty, we adopt the following procedure. For each MCMC iteration, based on the L samples of α, we set ζ to be their median and calculate $W(\alpha)$ numerically. If $W(\alpha)$ is less than a pre-specified threshold ξ, we continue to the next iteration. Otherwise, we will call in more training samples (Δ) and refine GPS following step 1 - 3 described above until the condition $W(\alpha) < \xi$ is fulfilled. The above workflow is summarized in Algorithm 4. The obtained samples are from the approximated posterior distribution, and we finally obtain the point estimator by taking the marginal mean of the samples after burn-in and obtain the credible interval based on sample quantiles. We call this new proposed

estimator as ABC-BD. The performance of this approach is compared with the MLE-BD in Section 2.1.4.

Algorithm 4 An GPS-ABC algorithm for estimation of mutation rate.

Inputs: $\mathbf{Y}$, S_0, Δ, θ, ϵ, ξ, $q(\cdot \mid \cdot)$, $\pi(\theta)$, m, N, $k(\cdot,\cdot)$, the simulator.
Step 1. Obtain the features from $\mathbf{Y}$ to get $\bar{\mathbf{d}}_y$.
Step 2. Determine design points Θ of size S_0. Generate initial training points $\mathbf{X}$ from the simulator at each grid point in Θ. Calculate feature values for each $\mathbf{X}$ to get $\{(\theta_i, \bar{d}_{x,i}), i = 1, \ldots, S_0\}$. Estimate the GP surrogate model.
Step 3. Run the MCMC iteration.
for $i = 1$ to N **do**
  Propose θ^* from $q(\theta^* \mid \theta)$.
  while $W(\alpha) > \xi$ **do**
    Calculate the mean and covariance for (θ^*, θ) following Equations (2.3)-(2.4)
    Generate L samples of $\mu_{\theta^*}^{(l)}$ and $\mu_{\theta}^{(l)}$ and calculate $\{\alpha^{(l)}, l = 1, \ldots, L\}$ using Equation (2.5).
    Set $\zeta = \text{median}(\{\alpha^{(l)}, l = 1, \ldots, L\})$ and calculate the probability of making a wrong decision $W(\alpha)$.
    if $W(\alpha) > \xi$ **then**
      Add Δ grid points to Θ, generate new training data at each newly added grid point. Add these points to the existing training data. Update the surrogate model.
    end if
  end while
  if $u < \zeta$ **then**
    Set $\theta = \theta^*$ and save θ.
  else Save θ.
  end if
end for

Comparing with the traditional ABC approaches, the GPS-ABC approach has several advantages: (1) Gaussian Process regression, as a widely used nonlinear modeling technique, requires only a small number of initial training samples [32]. Even when the parameter is of relatively high dimension, by adopting some design of experiment (DOE) technique such as Latin hypercube design (LHD) [40], we could still achieve satisfied performance with moderate number of training samples. This makes the method scalable to high dimensional estimation problems. (2) Instead of using the generated samples only once, the surrogte

model is able to "memorize" the generated samples in the model and mimic the behavior of the simulator. We only call in more samples to refine the model when its performance is not satisfied. Thus, GPS-ABC can dramatically reduce calls from the real simulator and saves us a great amount of computational resources. (3) As generating samples from the surrogate model is computationally efficient, we are able to quantify the probability of making wrong decision more easily and use the result to update the surrogate model, which is known as "uncertainty control".

2.1.3 Generalized birth-death process model and the estimator

We now consider generalizing the constant mutation rate p to a more general function $p(t)$. The simplest extension is to assume that $p(t)$ is a piecewise constant function, e.g., $p(t) = p_1 \mathbb{1}_{0<t\leq\tau} + p_2 \mathbb{1}_{\tau<t\leq t_k}$, where t_k is the known checking time. Under this setup, the underlying parameters include the changing time point τ, the mutation rate before τ denoted by p_1 and the mutation rate after τ denoted by p_2. We hope to estimate all three parameters simultaneously. It is hard to find the point mass function of the non-mutant cell population size in this scenario, which is needed to derive MLE. However, it is easy to incorporate this modification in the simulation model. We only need to replace Algorithm 1 with Algorithm 5 for the simulation model, and follow the procedure described in Algorithm 4. This gives an approximated Bayesian estimator for which we call ABC-GBD. There are only two things to adjust in order to deal with the scalability brought in by the higher dimension of the parameter space (from 1-dimension to 3-dimension): (i) the feature to summarize the two outputs and (ii) the way to select grid points.

Recall that, in the one-dimensional case discussed above, we have used $d_y = \sqrt{Z_t/N_t}$ as the feature to summarize information based on the two observed numbers: the number of overall

Algorithm 5 The generalized birth-death process model with piecewise constant $p(t)$

Inputs: a, p_1, p_2,τ,t
Step 1. Calculate the lifetime T of the first generation, which follows exponential distribution with rate $1/a$. Initialize culture size N_t by the number of cells with T greater than t and number of mutant cells $X_t = 0$
Step 2. Count X_t and N_t up to time t
while $T < t$ **do**
Update lifetime T by $T + T_new$, $T_new \sim Exp(1/a)$
Cells with $T < t$ will mutate with probability p_1
if C **then**omes from a mutant cell
Mutate at probability 1;
elseWhen $T < \tau$, mutate at rate p_1. Otherwise, mutate at rate p_2; Mark if mutated.
end if
Update $N_t = N_t$ + number of cells with $T > t$, $X_t = X_t$ + number of mutant cells
end while

cells N_t and the number of mutant cells Z_t. In order to make the regression function that maps the parameters to the feature smooth and non-flat, which helps reduce the number of training samples required for GPS, we use $d_y = \sqrt[4]{Z_t/N_t}$ as the feature in the 3-dimensional setup. and set the parameters in log scale, i.e, $\theta = (log(p_1), log(p_2), log(\tau))$.

With a constant mutation rate, we have 1-dimensional parameter space. The calculation is fairly easy even when the training samples are generated from a relatively dense grid on the domain. In more general case, however, we need to deal with a 3-dimensional parameter space. Assuming we have $S_0 = 50$ for the 1-dimensional case, we need to use $S_0 = 50^3 = 25000$ initial samples for the 3-dimensional case to achieve the same sampling density. In order to fit GPS with accuracy with less training samples, we adopt Latin-Hyper cube design (LHD) [40] to determine the training grid points. Our simulation shows that, if we use 50 grid points as initial samples to estimate GPS model for the 1-dimensional case, we only need 150 design points for the 3-dimensional case. Similarly, when refining the GPS model, we also use LHD to add training points.

With these two adjustments, we can easily find the estimator of θ using GPS-ABC algorithm

(Algorithm 4). Detailed steps for GPS-GBD are descrbied in Algorithm 5.

2.1.4 Simulation study and real data example

For the purpose of validation and assessment, we perform two simulation studies and apply the new estimators ABC-BD and ABC-GBD on one real data example. In Simulation 1, we evaluate the performance of the proposed estimator ABC-BD and MLE-BD for estimating the constant mutation rate p with different number of parallel cultures and different levels of mutation rates. Simulation 2 shows how the joint posterior distribution obtained by ABC-BD matches the truth under the more generalized birth-death process setup. Finally, the large scale case ($t = 19$, p on 10^{-8} scale) is illustrated by applying both ABC-BD and ABC-GBD estimators to a real data example in a bacteria fluctuation experiment. In this example, a saturated broth culture of *Staphylococcus aureus* was dispensed into 30 independent cultures, each started from 90 bacterium cells, and the average total number of cells of each culture is 1.9×10^8. Detailed experimental procedure can be found in Demerec [15] and Zheng [78]. For convenience, we list the experimental data in Table 2.1

Culture no.	1	2	3	4	5	6	7	8	9	10
Colony of mutants	33	18	839	47	13	126	48	80	9	71
No. of bacteria ($\times 10^8$)	1.83	1.79	1.82	1.79	2.02	2.05	1.76	1.85	2.06	2.02
Culture no.	11	12	13	14	15	16	17	18	19	20
Colony of mutants	196	66	28	17	27	37	126	33	12	44
Culture no.	21	22	23	24	25	26	27	28	29	30
Colony of mutants	28	67	730	168	44	50	583	23	17	24

Table 2.1: Data from penicillin-resistant fluctuation experiment[15]

Simulation 1: Comparison to MLE-BD

Using the forward simulation model shown in Algorithm 1, we generated data on a small scale

(checking time $t = 9$) for $J = 5, 10, 15, 20, 15$ parallel cultures. We set the true parameter $\theta = log(p)$ to be all integers ranging from -10 to -2, which corresponds to the true mutation probabilities ranging from 4.54×10^{-5} to 0.135. The domain of θ was set to be [-10.5, -1.5] on which the training samples were drawn. The size of initial design points was 50 and they were equally spaced in the domain of θ. At each design point, there were 10 replications. The prior of θ was a truncated normal distribution, with the mean set to be the logarithm of the MOM estimator and s.d. set to be 100, and bounded by the domain. The pre-specified threshold for making wrong decision is 0.25. The transition step was tuned to make the acceptance rate between $20\% \sim 40\%$. We obtained 30000 MCMC samples and used the first 10000 samples as burn-in. We then compared the accuracy of ABC-BD with MLE-BD by mean squared error (MSE) based on 100 simulations. Results are shown in Figure 2.3. From Figure 2.3, we observe that: (1) More cultures seem to bring very limited benefits in improving the accuracy of estimation. When the true mutation probability is large (e.g., 0.135), more cultures seem to lead to lower MSE for both methods. However, when the mutation probability is small, we cannot see any advantages of observing more cultures. (2) For the small scale example, MSE increases as the true mutation probability increases. When the mutation probability is 0.135, the MSE is at the 10^{-4} level, whereas it is around 10^{-8} when the true value is 4.54×10^{-5}. The estimators are pretty accurate based on the simulation study. (3) At a typical culture number of 15, the MSE of ABC-BD is generally less than MLE-BD, especially when the mutation probability is small.

Simulation 2: Evaluation of ABC-GBD estimator

In the second simulation, we illustrate the performance of the ABC-GBD estimator under a generalized birth-death process model. To simulate data, we set $p_1 = 3.4 \times 10^{-4}, p_2 = 0.018$ and $\tau = 4$. We consider a small scale case with checking time $t = 9$. There were 100

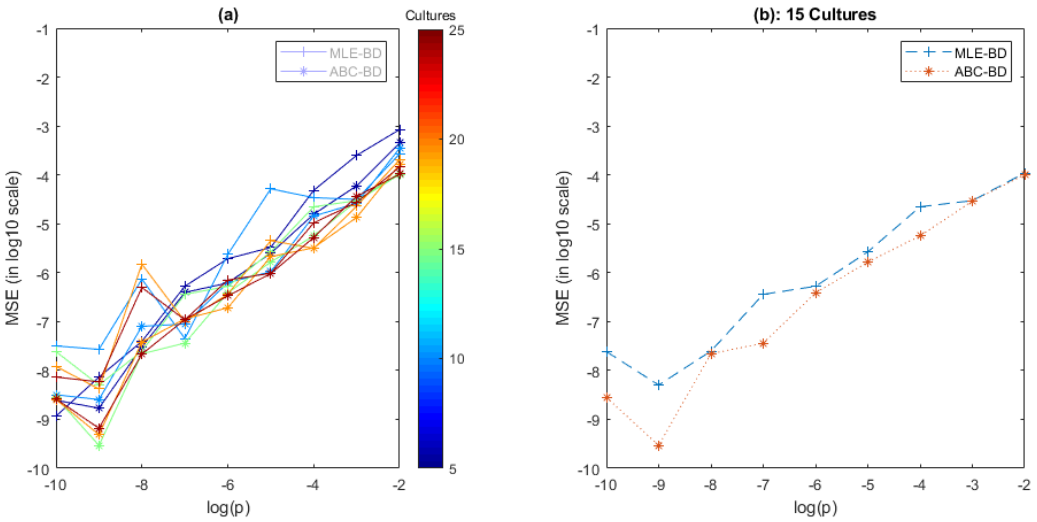

Figure 2.3: Small-scale simulation results of ABC-BD estimator. When t=9 and $log(p)$ at different levels of from -10 to -2: (a) shows mean squared error (MSE) in log scale for different number of parallel cultures as of different colors. Darkest Red stands for 25 cultures, and darkest blue stands for 5 cultures. (b) shows for a typical culture number of 15, the MSE of two methods at different levels of $log(p)$. Blue dashed line stands for the MLE-BD and red dotted line stands for ABC-BD estimator

independent parallel cultures. To setup the ABC estimation, we set the priors as follows: $\theta_1 = log(p_1)$, $\theta_1 \sim N(log(MOM), 20)$, $\theta_2 = log(p_2)$, $\theta_2 \sim N(log(MOM), 20)$ and $\theta_3 = log(\tau)$, $\theta_3 \sim N(1.5, 100)$. The domain of $\Theta = (\theta_1, \theta_2, \theta_3)$ was a cube bounded by $[-9, -1] \times [-9, -1] \times [-0.1, 2.2]$. We have $S_0 = 150$ initial design points based on LHD, obtained using the function "lhsdesign" in MATLAB. The divergence level ϵ was set to be $1e - 8$. The pre-specified threshold for making a wrong decision is 0.3. In each iteration, if the unconditional probability of making a wrong decision was larger than 0.3, we call in 10 more design points. We generated 50000 posterior samples in total and used the first 10000 samples as burn-in. Based on the retained 40000 samples, we achieve the result shown in Figure 2.4. Figure 2.4 provides the joint posterior distribution of every two parameters with the marginal distributions shown along the axises. These plots show that the ABC estimator is able to demonstrate multi-mode behavior of the parameters which cannot be handled by

	p_1	p_2	τ
Meaning	Stage I mutation prob.	Stage II mutation prob.	Changing time
Domain	$[1.23 \times 10^{-4}, 0.368]$	$[1.23 \times 10^{-4}, 0.368]$	[0.9, 9]
True Value	3.4×10^{-4}	0.018	4
Post.mean	5.6×10^{-3}	0.011	3.2
90% credible interval	$[2 \times 10^{-4}$,0.09]	$[5 \times 10^{-4}, 0.047]$	[1 8]

Table 2.2: The posterior estimation for the three unknown parameters in Generalized birth-death process model

the traditional methods. It provides an accurate estimation of the second stage mutation probability. For the other two parameters, the joint distribution forms a v-shape if the changing time is smaller than 2. This means that the mutation probability changes in the very beginning, we cannot tell how large the first stage mutation probability is unless it is large enough (e.g at 10^{-1} level). It is not a surprise because that if p_1 and τ are both low in value, there's only few cells mutating in the first stage, and thus we will lack information to make inference of them. The marginal mean and 90% credible interval are given in Table 2.2, which tells the similar story as Figure 2.4: The marginal posterior means are close to the true values. The estimation of p_2 is more precise than the ones of p_1 and τ

An example on real data analysis

We applied ABC-BD and ABC-GBD on a published data set from a penicillin resistant fluctuation experiment conducted by Demerec [15] and compared the resulting point estimate with that obtained by MLE-BD. Since the overall sizes are only available for the first 10 cultures, we used the average culture size 1.9×10^8 for the rest 20 cultures. Based on the first 10 cultures, MLE-BD estimator reports mutation probability estimate as 1.82×10^{-8}. For the next 20 cultures, the value is 1.65×10^{-8}. For the merged dataset, the value becomes 1.71×10^{-8}.

To apply ABC-BD, we first set up the simulation model. Notice that this is a large scale

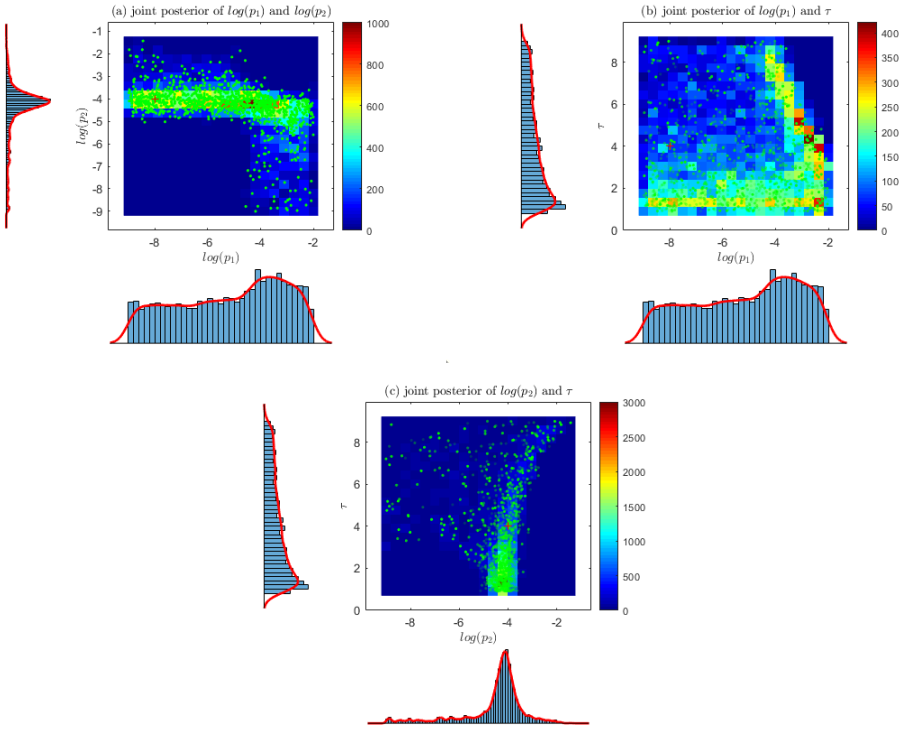

Figure 2.4: Results for the ABC-GBD estimator. The heatmaps of the 2-d kernel density estimations for each pair of parameters: (a) $log(p_2)$ vs $log(p_1)$; (b) τ vs $log(p_1)$; (c) τ vs. $log(p_2)$. The green dots on the heatmaps are the scatter plots fo the posterior samples. The red cross symbol marks the true parameter values. The histograms on the left and bottom of each heatmap are the marginal distributions for each parameter.

case with the culture size at 10^8 level. In order to simulate a group of cells with the same culture size, we need to set up the simulation model with larger checking time t than the small scale case discussed in the simulation studies. The checking time is set to be 19, and $a = 1$ as suggested by Wu and Zhu [74]. To setup the estimator, the prior of $\theta = log(p)$ is set to be $N(log(MOM), 10)$. The divergence level ϵ is 1e-8. There are 20 initial design points, with 4 replications at each design point. The threshold ξ of making error is 0.25. If the unconditional expected probability of making wrong decision is greater than 0.25, we will call in 5 more design points. Design points are equally located on the domain of θ, [-20,-10]. The MH sampler was tuned to obtain an acceptance rate of 51%. We obtained

	p_1	p_2	τ
Meaning	Stage I mutation prob.	Stage II mutation prob.	Changing time
Domain	$[2.06 \times 10^{-9}, 4.54 \times 10^{-5}]$	$[2.06 \times 10^{-9}, 4.54 \times 10^{-5}]$	[0.9,19.1]
Post.mean	1.59×10^{-7}	4.29×10^{-9}	3.4505
90% credible interval	$[3.36 \times 10^{-9}, 1.32 \times 10^{-5}]$	$[2.19 \times 10^{-9}, 1.18 \times 10^{-8}]$	[0.96,13.88]

Table 2.3: The ABC-GBD estimation based on the real data

10000 MCMC samples and leave the first 8000 as burn-in. The retained 2000 samples give as the posterior mean of 2.81×10^{-8} for the first 10 cultures, 3.24×10^{-8} for the rest 20 cultures and 2.74×10^{-8} for the merged dataset. The results are very close to the MLE-BD estimator with slightly larger values.

We set the simulation model same as above in the ABC-GBD case. For the estimator setup, we let the priors be $\theta_1 = log(p_1)$, $\theta_1 \sim N(log(MOM), 20)$, $\theta_2 = log(p_2)$, $\theta_2 \sim N(log(MOM), 20)$ and $\theta_3 = log(\tau)$, $\theta_3 \sim N(2.3, 100)$. There are 150 initial design points, on each of which, with 3 replications on each point. The threshold ξ of making error is set to be 0.3. We will call in 10 more design points if the expected probability of making error is larger than ξ. Design points are determined by LHD. We obtained the posterior estimation shown in Table 2.3. The posterior samples are further plotted as paired joint distribution in Figure 2.5.

From Figure 2.5, we observe that the posterior distributions of the parameters demonstrate skewed, multi-modality shapes. In particular, the marginal distribution of the changing time and log of probability of Stage II is skewed to the right, and the log of probability of Stage I demonstrates two modes, one near -18 and the other near -13. Furthermore, the 90% credible intervals for p_1 and t are fairly wide. Wide credible intervals indicate high uncertainty in the point estimates. These results are not a surprise, because they reflect several characteristics of the generalized birth-death process model: The multi-modal behavior of the posterior distribution reflects the non-identifiability nature of the inverse-

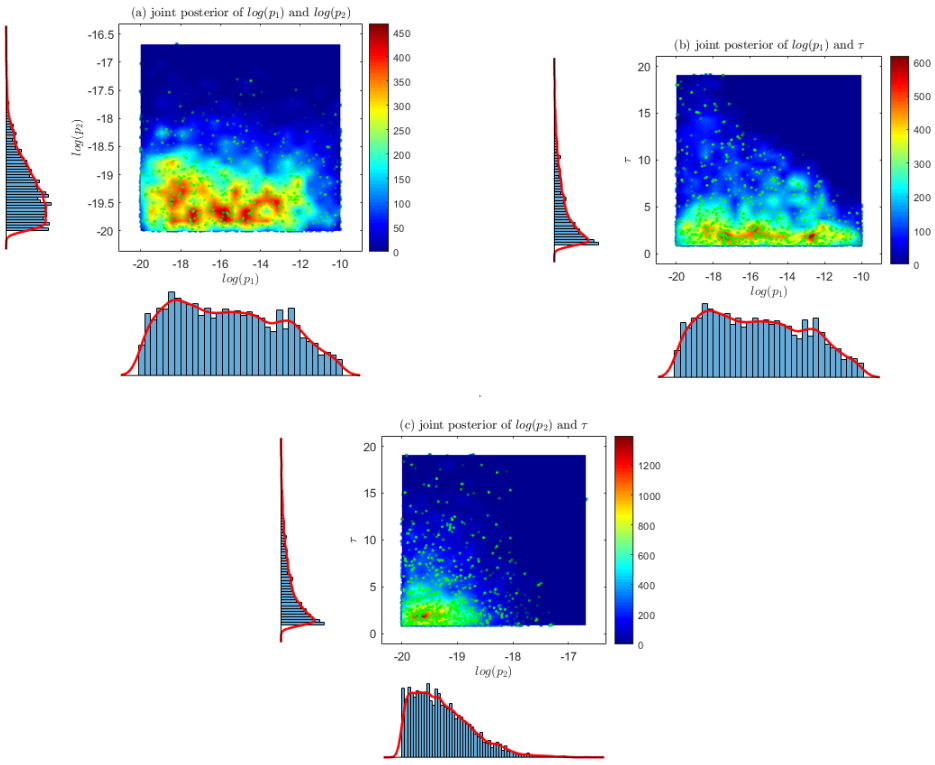

Figure 2.5: [Real data analysis results using the ABC-GBD estimator. Based on the real data shown in Table 2.1, we obtain the heat-maps of the 2-d kernel density estimations for each pair of parameters: (a) $log(p_2)$ vs $log(p_1)$; (b) τ vs $log(p_1)$; (c) τ vs. $log(p_2)$. The green dots on the heat-maps are the the posterior samples. The red cross symbols mark the true parameter values. The histograms on the left and bottom of each heat-map are the marginal distributions for each parameter.

problem, i.e., different combinations of the p_1, p_2 and τ could result in similar numbers of overall cells and mutant cells . Therefore, the solution to the inverse-problem is not unique. Despite these challenges, our proposed method still provide a comprehensive view for the distributions of the underlying parameters under a moderate number of samples—a result that is intractable if using any other existing statistical approaches. These results demonstrate the promise of solving ill-posed inverse-problems.

2.2 Estimating Parameters in Complex Systems with Functional Outputs—A Wavelet-based Approximate Bayesian Computation Approach

2.2.1 Introduction

Functional data, such as signals, surfaces, and images, are frequently encountered in many scientific disciplines. The increased prevalence of such data promotes the development of *functional data analysis* [17, 22, 53, 68]. While considerable efforts have been made to the preprocessing [52, 63], estimation [54, 76, 77], and regression analysis [10, 12, 45, 57, 81] of functional data, existing approaches primarily rely on linking functional observations with the unknown parameters via a likelihood or an objective function. Many applications, however, involve inferring parameters when such linkage is implicit or difficult to specify. In this work, we consider a family of parameter estimation problems, in which the relationship between the functional observations and the unknown parameters cannot be specified explicitly.

Our research is motivated by the foliage-echo data arising from a sonar study. The goal is to infer the parameters of targets (i.e., foliages) based on the echo signals. The foliage-echo example represents a general class of physical systems with functional data outputs. These systems have the following characteristics: (1) Due to the complexity of the physical rules, the parameter estimation is a difficult inverse-problem which may be ill-posed. Analytical or numerical solutions may be hard to find, and the solutions may not be unique. (2) One can numerically simulate data from the physical system, even though the simulation can be computationally intensive. (3) The data-generation procedure of the physical system is random. It produces random functional outputs for a given set of parameters. (4) It is

often difficult to explicitly link the functional outputs with the underlying parameters via a likelihood or an objective function. (5) The functional outputs are high-dimensional.

The difficulty of explicitly linking functional data with the underlying parameters makes most statistical approaches non-applicable. It is possible to assume a statistical model, for example, a regression that treats the unknown parameters as the responses and the functional data as predictors. With such an assumption, the prediction of the unknown parameters can be achieved through a training-testing procedure, i.e., first training the model using the training data, and then predicting the unknown parameters using the test data. However, such statistical models completely ignore the underlying data-generating mechanism, thus do not reflect the true data-parameter relationship. In certain circumstances, for example, when two sets of parameters result in the same functional output, a simple statistical model like a linear regression often fails to provide even a reasonable estimation.

To tackle the parameter estimation challenges in complex systems, we propose a wavelet-based approximate Bayesian computation (wABC) approach. The proposed approach inherits the "likelihood-free" property of the traditional approximate Bayesian computation (ABC) [38, 64]. We propose to first apply wavelet decomposition and compression to reduce the dimension and decorrelate the functional data and then adopt the GPS-ABC algorithm introduced in Section 2.1.2 with the remained wavelet coefficients.

To our knowledge, the proposed approach is the first that estimates parameters in complex systems based on high-dimensional functional outputs. It is generally applicable to various physical, chemical, and biological systems that facilitate numerical simulations. Compared with existing functional data analytical tools, our approach has the following advantages: (i) It is likelihood-free. It takes full advantages of the mathematical/physical rules that connect data with the parameter. (ii) It can characterize various linear or nonlinear data-parameter relationships. (iii) It produces the joint posterior distribution of the parameters with vari-

ous multi-modality and shape structures. (iv) It is scalable to high-dimensional functional outputs and expensive simulations. Our results for the foliage-echo data demonstrate the effectiveness of the proposed method in estimating parameters.

In the next several sections, The wavelet representation and compression of functional data is introduced in Section 2.2.2. Based on it, we propose our method wABC in Section 2.2.3 and apply the method for the foliage-echo simulation model in Section 2.2.4

2.2.2 Wavelet representation and compression of functional data

While the idea of ABC is straightforward to follow, it can be inefficient due to a number of assumptions and approximations that may not be easily satisfied. One assumption is the existence of a sufficient statistics for the parameters of interest. Given a random sample $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$, the determination of a sufficient statistic $S(\mathbf{Y})$ for θ is often difficult without knowing the distribution of Y_i. Although one can always choose the data itself as the sufficient statistic, doing so only makes the specification of the distance measure $\rho(\cdot, \cdot)$ extremely difficult (because the dimension of Y is high). This issue is particularly severe for high dimensional vectors and functional data. In our foliage-echo example, an echo envelope is of dimension $24,000$. Therefore the data Y can be written as a matrix of size n by $24,000$. Given that the relationship between the data and the parameters is implicit, determining sufficient statistics for $(\theta_1, \theta_2, \theta_3)$ given $\mathbf{Y}$ is practically intractable.

To facilitate the efficient performance of ABC for high-dimensional functional data, we adopt a strategy that achieves de-correlation and dimension reduction so that functional observations can be parsimoniously represented in a much lower-dimensional setting. In particular, we represent the functional data by a multi-scale wavelet basis. Given a set of multi-scale wavelet basis functions $\{\psi_{jk}; j = 1, \ldots, J,\ k = 1, \ldots, K_j\}$ and a scale function

(the father wavelet) $\{\psi_{0k}; k = 1, \ldots, K_0\}$, we can expand a functional observation $Y(t)$ by $Y(t) = \sum_{j=0}^{J} \sum_{k=1}^{K_j} d_{jk}\psi_{jk}(t)$. Here, d_{jk} is the wavelet coefficient at scale j and location k. For functional data measured on an equally spaced grid, this representation is *lossless*, i.e., providing an exact representation of the original data. Therefore, $\{d_{jk}\}$ contain the same amount of information as $Y(t)$ thus can be treated as a sufficient statistic for θ. We can denote the sufficient statistics of Y as $S(Y) = $, where $ = (d_{ijk})$ is a matrix of size n by K, with $K = \sum_{j=0}^{J} K_j$. In general, the wavelet transformation is not the only option. It is possible to construct lossless transforms with other basis functions (e.g. Spline or Fourier bases), or construct an approximately-lossless transformation with a basis $\{B_k(t), k = 1, \ldots, K\}$ that satisfies $|Y(t) - \sum_{k=1}^{K} d_k B_k(t)| < \delta$ for all t and a small δ.

The wavelet representation has two advantages: the coefficients $\{d_{jk}\}$ are sparse (meaning that most coefficients are zero or close-to-zero) and they are approximately uncorrelated. These properties bring several conveniences to the specification of the distance measure in ABC. First, since components in $\{d_{jk}\}$ are approximately uncorrelated, the conditional distribution $\pi_\epsilon(S(\mathbf{Y}) \mid S(\mathbf{X}))$ can be specified following Equation (1.1), i.e., assuming that components of $S(\mathbf{Y})$ (or $S(\mathbf{X})$) are mutually independent of each other. Second, the sparsity of the wavelet coefficients makes the wavelet compression feasible.

Wavelet Compression Like many high-dimensional problems, representing the data in a much lower dimensional space brings tremendous convenience to data storage and processing. This is also true in the ABC context. Let $\mathbf{D} = (d_{ijk})$ denote the n by K matrix of wavelet coefficients, with the ith row corresponding to the wavelet coefficients of the ith functional observation. Since $\mathbf{D}$ is sparse, many components of $\mathbf{D}$ are zero or close-to-zero, therefore does not contain essential information about the parameter. Wavelet compression helps remove zero or close-to-zero components while retaining the large components. The compressed matrix, denoted by $\widetilde{\mathbf{D}}$, is nearly lossless, thus can be used as an approximately-

sufficient statistic for θ. To compress $\mathbf{D}$, we retain K_1 columns of $\mathbf{D}$ so that the total amount of energy retained is greater than or equal to a threshold δ_1 (e.g., $\delta_1 = 0.999$) for each function. Here, the total energy for a function $Y_i(t)$ is defined by $\sum_{(j,k)\in C_1} d_{ijk}^2 / \sum_{(j,k)} d_{ijk}^2$, where C_1 is the set of scale and location indices that correspond to columns retained.

2.2.3 Wavelet-based ABC(wABC) approach

We explain the GPS model in the context of the foliage-echo example by just generalizing the previous quantity calculations of 1 dimensional case to the multiple dimensional case to allow J features. Suppose that J columns of are retained after wavelet compression. Denote by$\widetilde{}_y = (\overset{1}{}_y, \ldots, \overset{J}{}_y)$ the n by J matrix of wavelet coefficients after compression, in which each $\overset{j}{}_y$ is a n by 1 vector. In the foliage-echo example, since $Y = \{Y_1, \ldots, Y_n\}$ contains n echo envelope signals, each $\overset{j}{}_y$ is an n by 1 vector. The randomness in the leaf location, orientation, and radius causes random fluctuations in the the n samples. These fluctuations reflect the scene-specific information, i.e., exact locations, orientations, and radii of leaves in a scene, which is not relevant to the population parameters $(\theta_1, \theta_2, \theta_3)$. Therefore, we remove the random fluctuation by averaging each $\overset{j}{}_y$ across its n entries, resulting in a scalar $\bar{d}_y^j$. Denote the averaged wavelet coefficients by$\bar{}_y = (\bar{d}_y^1, \ldots, \bar{d}_y^J)^T$. We will use $S(Y) =\bar{}_y$ in the analysis of foliage-sonar data. Since the wavelet coefficients in$\bar{}_y$ are approximately independent of each other, we will calculate the likelihood $\pi_\epsilon(\bar{d}_y^j \mid \theta)$ for each j independently. We then assume that

$$\bar{d}_y^j = \bar{d}_x^j + e^j, \quad e^j \sim N(0, \epsilon^2). \tag{2.6}$$

Here, $\bar{d}_x^j$ is the jth averaged wavelet coefficients based on the simulated samples $X =$

$\{X_1, \ldots, X_m\}$. We further assume that

$$\bar{d}_x^j = f_j(\theta) + r_j, \quad f_j(\theta) \sim GP(0, k_j(\theta, \theta^*)), \quad r_j \sim N(0, \sigma_j^2), \tag{2.7}$$

where $f_j(\theta)$ is an unknown Gaussian process (GP) with mean zero and a pre-specified covariance kernel $k_j(\theta, \theta^*)$. For example, a commonly used covariance kernel is the squared exponential kernel $k_j(\theta, \theta^*) = \phi_j^2 \exp\{-||\theta - \theta^*||^2/(2\tau_j^2)\}$.

Specifically, we calculate $\pi_\epsilon(\bar{d}_y^j \mid \theta)$ following a three-step procedure.

1. Produce a grid of values $\Theta = (\theta_1, \ldots, \theta_A)^T$ on the domain of θ, generate $X = \{X_1, \ldots, X_m\}$ at each grid point, perform wavelet decomposition and compression of X, and average the wavelets coefficients across the m samples. This results in a list of "input-output" pairs $\{(\theta_i, \bar{d}_{x,i}^j), i = 1, \ldots, A\}$, which will be treated as the training data for estimating the function $f_j(\theta)$.

2. Given a pair of values (θ^*, θ), we will calculate the GP predictive distribution on (θ^*, θ) using the conditional distribution, which gives $N(\mu^j_{(\theta^*,\theta)|\Theta}, \Sigma^j_{(\theta^*,\theta)|\Theta})$, where

$$\mu^j_{(\theta^*,\theta)|\Theta} = \begin{pmatrix} \mathbf{k}_{\theta^*,\Theta} \\ \mathbf{k}_{\theta,\Theta} \end{pmatrix} \left(\mathbf{K}_{\Theta,\Theta} + \sigma_j^2 \mathbf{I}\right)^{-1} \bar{\mathbf{d}}_x^j \,, \tag{2.8}$$

$$\Sigma^j_{(\theta^*,\theta)|\Theta} = \begin{pmatrix} k_{\theta^*,\theta^*} & k_{\theta^*,\theta} \\ k_{\theta,\theta^*} & k_{\theta,\theta} \end{pmatrix} - \begin{pmatrix} \mathbf{k}_{\theta^*,\Theta} \\ \mathbf{k}_{\theta,\Theta} \end{pmatrix} \left(\mathbf{K}_{\Theta,\Theta} + \sigma_j^2 \mathbf{I}\right)^{-1} \begin{pmatrix} \mathbf{k}_{\theta^*,\Theta} \\ \mathbf{k}_{\theta,\Theta} \end{pmatrix}^T . \tag{2.9}$$

   Here, $\bar{\mathbf{d}}_x^j = (\bar{d}^j_{x,1}, \ldots, \bar{d}^j_{x,S_0})^T$ is a S_0 by 1 vector of training pointsNote that here, we'll find the hyperparameters of GPS model for each j, meaning that the surrogate model is fitted independently for each dimension of selected features. In Figure 2.6, we compared the prediction performance of the GPS with the sample estimate obtained

from data directly sampled from the simulator. Figure 2.6 demonstrates that the GPS gives as accurate prediction as the sample estimates (which are based on 100 samples) using only 10 training locations on the support of θ_1.

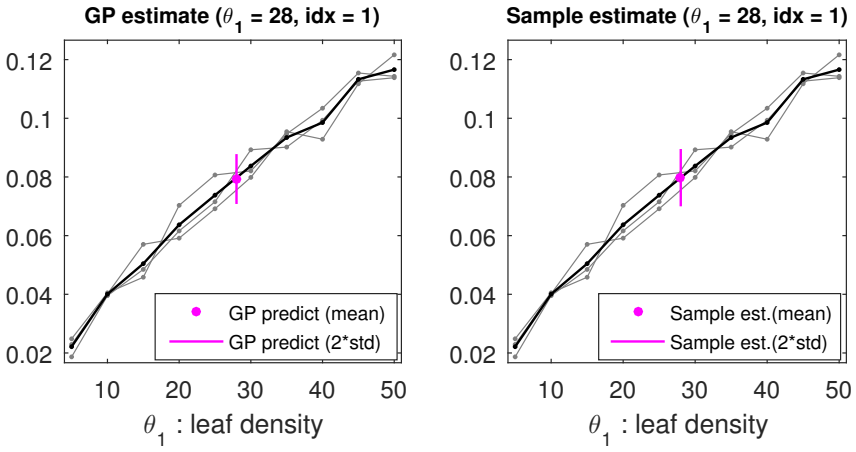

Figure 2.6: A one-dimensional demonstration of the GP prediction using the sonar-foliage simulator. Here, we have fixed $\theta_2 = 0.017$ and $\theta_3 = 45$, and treated θ_1 as the unknown parameter. Left panel: the gray lines are the first wavelet coefficient of $m = 3$ simulated echo envelopes at $A = 10$ grid points on the domain $[5, 50]$; the black lines are the average of the three gray lines; the magenta dot and line are the predictive mean and the confidence interval ($mean \pm 2\, std$) calculated using the GPS. Right panel: the gray lines and the black lines are the same as the left panel. The the magenta dot is the sample estimate of the mean, and the magenta bar is the confidence interval based on 100 echoes sampled directly from the simulator.

3. Based on the GPS, the likelihoods $\pi_\epsilon(\bar{d}_y^j \mid \theta^*)$ and $\pi_\epsilon(\bar{d}_y^j \mid \theta)$ can be approximated by $N(\bar{d}_y^j | \mu_{\theta^*}^{j,l}, \sigma_j^2 + \epsilon^2)$ and $N(\bar{d}_y^j | \mu_\theta^{j,l}, \sigma_j^2 + \epsilon^2)$ respectively, where $(\mu_{\theta^*}^{j,l}, \mu_\theta^{j,l})$ is a sample from $N(\mu^j_{(\theta^*,\theta)|\Theta}, \Sigma^j_{(\theta^*,\theta)|\Theta})$. The acceptance probability of the MCMC can be calculated by

$$\alpha^{(l)}(\theta^*|\theta) = \min\left\{1, \frac{\pi(\theta^*)\prod_{j=1}^J N(\bar{d}_y^j|\mu_{\theta^*}^{j,l}, \sigma_j^2+\epsilon^2) q(\theta|\theta^*)}{\pi(\theta)\prod_{j=1}^J N(\bar{d}_y^j|\mu_\theta^{j,l}, \sigma_j^2+\epsilon^2)\; q(\theta^*|\theta)}\right\}. \tag{2.10}$$

Finally, we just replace Equations 2.3, 2.4 and 2.5 with Equations 2.8 ,2.9 and 2.10 respectively in Algorithm 4 to achieve the estimation results. Note that for the numerical stability

of the algorithm and the convenience of setting parameters, we recommend to rescale the compressed data $\widetilde{\boldsymbol{D}}_y$ and the simulated features $\widetilde{x}$ using a common set of constants so that all values are in a similar scale (e.g., $[-1, 1]$). The scaling constants can be estimated from the observed data (e.g., using the minimum and maximum of $\overset{j}{y}$ for each j). Similarly, in the GPS calculation, we recommend to scale all parameters to a common range (e.g., $[0, 1]$).

Parameter settings. There are several parameters that need to be pre-specified in the wABC algorithm. The parameter ϵ in equation (2.10) is a small value that controls the expected discrepancy between simulated and observed data. We suggest to set a small value (e.g., 1e-4) for ϵ. In the GPS MCMC algorithm, it is possible to set $\epsilon = 0$ as done by [1]. The parameters ξ, m, N, S_0, and Δ can be tuned based on the computation speed and the acceptance rate of the MCMC algorithm. In general, the accuracy of the posterior estimation can be improved by increasing the sample size in data Y, reducing the threshold ξ for the probability of making an error in the MH sampler, increasing the size of the training grid for GPS, and increasing the number of training samples m at each GP training grid, as it may reduce the estimation of σ^j, which may make the prediction error smaller.

2.2.4 The Foliage-echo data analysis

We applied the proposed wABC approach to a set of foliage-echo data simulated from the sonar-foliage simulator. The data consists of $n = 100$ echo envelope signals sampled independently from the simulator under the true parameter $(\theta_1, \theta_2, \theta_3) = (30, 0.017, 45)$. We aim to solve the inverse-problem by estimating the three underlying parameters based on the 100 echo envelopes while assuming that the domains of the parameters are $\theta_1 \in [5, 50]$, $\theta_2 \in [.005, .05]$, and $\theta_3 \in [1\text{e-}4, 90]$.

We applied the wavelet transformation to each echo envelope using Daubechies wavelets with

the maximal number of vanishing moments being 12 (i.e., db12). The number of resolution levels is set to be $J = 20$, and the boundary extension mode is set to be periodic. The wavelet decomposition transforms each echo envelope from the time domain (with $24,000$ measurement points) to the wavelet domain (with $24,008$ wavelet coefficients). We further applied wavelet compression by retaining $\delta_1 = 0.999$ of the total energy. This reduces the dimension of the wavelet coefficients from $24,008$ to 992. We then applied MCMC with GPS using Algorithm 4. We adopted a random walk proposal by setting the proposal distribution $q(\theta^* \mid \theta)$ to be a truncated log-normal with a scale parameter 0.05. To train the GPS, we segmented the domain of $(\theta_1, \theta_2, \theta_3)$ using a $10 \times 10 \times 10$ equally-spaced grid. This gave a total of 1000 training points for the GPS. The number of repeated samples in X on each grid point was set to be $m = 3$. The ϵ parameter in the MCMC-ABC was set to be 1e-4 and the ξ parameter in the GPS procedure was set to be 0.3. These setups resulted in an acceptance rate of 35% in the MCMC MH sampler. We monitored the behavior of the posterior samples by checking the trace plots and the autocorrelation plots. We tested the convergence of the chains by calculating the Geweke's Z-statistics [18]. We ran $30,000$ MCMC iterations and took the first $10,000$ iterations as the burin period. Summary statistics of the parameter estimation, including the posterior means and the 95% credible intervals (CIs), are listed in Table 2.4. Table 2.4 shows that all three CIs cover the true values of the parameters.

We further summarized of the posterior distribution of the parameters using 1-d and 2-d marginal kernel density estimations. In Figure 2.7, we plot the heatmaps of the 2-d kernel density estimations for each pair of the parameters. The gray dots on the heatmaps are the scatter plots of the posterior samples (a total of $15,000$ samples after the burnin period). The white cross sign on the heatmaps mark the true values of the parameters. The histograms on the top and right-hand side of each heatmap show the marginal distributions of the parameters (superimposed by the 1-d density estimations). The red vertical bar in each

	θ_1	θ_2	θ_3
Meaning	density in 3-d	mean radius	mean orientation
Unit	counts per m^3	meter	degree
Domain	[5, 50]	[.005, .05]	[1e-4, 90]
True value	30	.017	45
Post. mean	28.45	.018	42.57
Post. CIs	[17.1, 41.9]	[.017, .026]	[10.1, 72.7]

Table 2.4: The posterior estimation for the three parameters in the foliage-echo data.

histogram indicates the location of the posterior mean, and the red dashed bars indicate the 95% credible interval.

From Figure 2.7, we observe that the posterior distributions of the parameters demonstrate skewed, multi-modality shapes. In particular, the marginal distribution of the leaf size is skewed to the right, and the leaf orientation demonstrates two modes, one near 10 degree and the other near 40 degree. Furthermore, the 95% CIs for the leaf density and the orientation are fairly wide. Wide CIs indicate high uncertainty in the point estimates. These results are not a surprise, because they reflect several characteristics of the foliage-echo simulation. First, the echo signals are highly stochastic—using 100 echoes samples to recover the statistical properties of the foliages is a challenging task. Second, the multi-modal behavior of the posterior distribution reflects the non-identifiability nature of the inverse-problem, i.e., different combinations of the leaf density, size, and orientation could result in similar reflection behavior of the sound wave. Therefore, the solution to the inverse-problem is not unique. Despite these challenges, our proposed wABC still provide a comprehensive view for the distributions of the underlying parameters under a moderate number of samples—a result that is intractable if using any other existing statistical approaches.

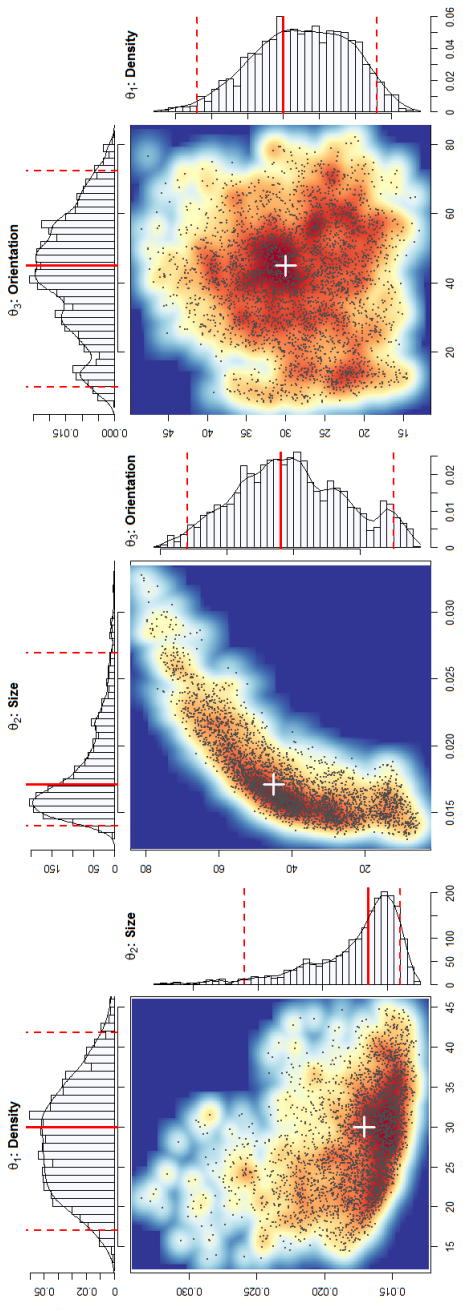

Figure 2.7: Results for foliage-echo data analysis. The heatmaps of the 2-d kernel density estimations for each pair of parameters. Left: θ_1 versus θ_2; middle: θ_2 versus θ_3; right: θ_3 versus θ_1. The gray dots on the heatmaps are the scatter plots of the posterior samples. The white cross symbol on the heatmaps marks the true parameter values. The histograms on the top and the right-hand side of each heatmap are the marginal distributions (histograms superimposed by the 1-d kernel density estimations) for each parameter. The red vertical bar in each histogram indicates the location of the posterior mean, and the red dashed bars indicate the 95% credible interval.

Chapter 3

Scalable functional region detection via multiple testing in basis-space

3.1 Introduction

With modern high-throughput technologies, scientists can now collect various types of high-dimensional data in functional form, with basic observational units being curves or surfaces measured over fine grids. Examples include longitudinal trajectories, engineering signals, brain images, and genomic measurements. While functional data often carry excessive amount of information, in many situations only a few local regions are relevant to the problem of interest. For example, in proteomics, peaks on a mass spectrometry curve represent abundance of certain proteins or peptides in a biological sample, which are the only useful information relevant to most proteomic analysis. In such cases, detecting relevant local regions plays important role in extracting useful information, searching for important biomarkers, and guiding decision-making.

Despite the pressing need for region detection, existing methodology is limited. Commonly used methods often rely on detecting components of a long vector obtained by discretizing functional data [7, 34, 62] or selecting among a pool of features extracted from pre-defined local regions [6, 35, 75]. While such ad hoc methods may be useful, naively transforming functional data to vectors can be problematic and result in either missed regions or false dis-

coveries. In particular, discretization-based methods can lead to results that are susceptible to different discretizations or refinement of measurement grids, and feature-based methods can be sensitive to different ways of feature extraction. Additionally, as the resolution of functional data increases, these methods often suffer from computation bottleneck caused by curse of dimensionality [44].

In functional data analysis literature, relevant regions may be detected in a functional regression framework by constructing confidence bands for the regression coefficient function [9, 11, 37] or controlling overall FDR. For example, Meyer et al. [43] proposed to detect regions on regression coefficient functions that are significantly nonzero using the Simultaneous Band Scores (SimBaS) method and [46] proposed an approach to detecting regions on regression coefficient function that are greater than a pre-specified threshold while controlling the expected Bayesian false discovery rate (Bayesian FDR). These approaches have been used in Bayesian functional mixed models and functional response regression frameworks [81, 82, 84] for detecting regions with significant contrast effects between groups. For methods that do not have established confidence bands or do not produce posterior samples, these approaches can also be applied by generating Bootstrap samples of the regression coefficient estimation [11, 14]. Despite their effectiveness, methods that are based on confidence bands are often restricted to specific regression models and the corresponding asymptotic properties can be hard to establish. Additionally, approaches such as SimBaS rely on controlling FWER across a discretized grid which suffers from reduced power as the number of grid points becomes extremely large.

Besides methods based on confidence bands, detection of relevant regions can also be achieved by encouraging zero sub-regions on the regression coefficient function in functional regression frameworks. For example, James et al. [26] used a p-dimensional basis to approximate the regression coefficient function in a functional linear regression framework and applied a

lasso-type penalty to the discretized approximation of the dth derivative of the regression coefficient function. When d equals to zero, this gives a sparse estimation of the regression coefficient with zero sub-regions. Following a similar idea, Zhou et al. [80] also considered functional linear regression and aimed to encourage zero-values of coefficient function at sub-regions. They proposed a two-stage estimator by combining a Dantzig selector and a group SCAD penalty. Alternatively, Koltchinskii and Minsker [29] focused on L_1-norm penalization, a natural extension of Lasso, while assuming the regression coefficient function to be sparse, and Lin et al. [33] proposed a functional generalization of the SCAD penalty named fSCAD, and combined fSCAD with smoothing splines to achieve smoothing and locally sparse estimation. While achieves region detection, these existing methods only focus on functional linear models which characterizes association between a functional predictor and a continuous response variable. Extension to more complicated situations such as categorical response variables is not available. Furthermore, the performance of these methods has only been evaluated using data with moderate size, and the scalability to extremely high-dimensional cases such as brain images has not been studied.

Alternatively, functional region detection can be achieved by hypothesis testing. Existing literature on functional testing has focused on global tests between groups, i.e., testing equality of distributions [8, 21, 27, 50], population means [19, 56, 60] or covariance operators [20] across groups of functions. These global tests address whether there exists difference between groups but cannot identify regions on which the differences appear. Notable works that achieve region detection include Cox and Lee [13], Vsevolozhskaya et al. [67], Pini and Vantini [48] and Pini and Vantini [49]. Cox and Lee [13] considered point-wise testing using Westfall-Young randomization adjustment and proved the grid-refinement property for the corrected point-wise p-values. Significant local regions may be detected by thresholding the corrected point-wise p-values. Vsevolozhskaya et al. [67] proposed to perform functional

testings on some pre-defined sub-intervals and correct the interval-wise p-values using multiplicity adjustment. Pini and Vantini [48] projected functional data on a finite dimensional basis and performed a family of tests that includes all consecutive combinations of the basis components. Pini and Vantini [49] defined an unadjusted and an adjusted p-value function based on functional tests on arbitrary intervals. While these approaches are effective for region detection in various contexts, they are limited in several aspects. For example, the approach of Cox and Lee [13] requires performing point-wise testing on a grid, which can be computationally challenging for data with high number of grid points. Vsevolozhskaya et al. [67]'s method heavily depends on pre-defined intervals. Pini and Vantini [48]'s method requires a large family of tests which may increase exponentially with the number of basis functions, and Pini and Vantini [49]'s method requires functional tests on arbitrary (infinitely many) intervals which is numerically challenging. Additionally, these existing approaches have only been accessed using 1-dimensional functional data of moderate size. Their scalability to high-dimensional cases has not been studied.

In this chapter, we propose a testing procedure to detect regions on functional data that are significantly different across groups. The proposed procedure adopts compactly supported, potentially multi-resolution basis to capture local features of functional data, and achieves scalable region detection by performing a p-value-guided compression followed by simultaneous testing in basis-space. Compared with existing methods, the proposed procedure offers several unique features and advantages: (1) It facilitates scalable inference through parsimonious data representation, compression, and parallel computation on manycore CPUs or multicore GPUs, thus is scalable to extremely high-dimensional data. (2) The p-value guided compression offers improved power compared to point-wise testing in data domain. (3)The significant regions are detected and conveniently visualized after inverse-transforming significant components back to the original data space, enabling straightforward interpretation.

(4) The basis-space testing retains nice theoretical properties such as strong control of FWER/FDR and asymptotic optimality if using the Westfall-Young randomization adjustment. (5) It offers a flexible framework that can be combined with various parametric/nonparametric tests and multiplicity adjustments. We will evaluate the performance of the proposed procedure using two simulation studies. The procedure will be applied to two real datasets: a 1-dimensional fluorescence spectroscopy data and a 3-dimensional neuroimaging data. The latter is of extremely high-dimensionality with number of measurement points in the order of $O(10^7)$.

The outline of this chapter is as follows. Section 3.2 introduces the testing problem and the proposed basis-space testing procedure. Section 3.4 describes the related theoretical properties. Simulation results are presented in Section 3.5. Real data applications on fluorescence spectroscopy data (described in Section 1.2.3) and neuroimaging data (described in Section 1.2.4) are demonstrated in Section 3.6.

3.2 A Basis-Space Testing Procedure for Region Detection

3.2.1 Problem setup

Let $y_{ij}(x), i = 1, 2, \ldots, g; j = 1, \ldots n_i$ be random functional variables observed on domain $\mathcal{T}$. We consider testing linear relationship among the group means, i.e., for $x \in \mathcal{T}$,

$$\begin{aligned} H_0 : \mathbf{A}\,\boldsymbol{u}(x) &= 0 \\ H_a : \mathbf{A}\,\boldsymbol{u}(x) &\neq 0 \end{aligned} \tag{3.1}$$

where $\boldsymbol{u}(x) = (\mu_1(x), \ldots, \mu_g(x))^T$ contains the mean functions of the g groups, with $\mu_i(x) = E\{y_{ij}(x)\}$; and $\mathbf{A}$ is a constant matrix that constitutes different types of contrast effects. For example, a special case of the above construction is the two-sample test

$$H_0 : \mu_1(x) = \mu_2(x)$$
$$H_a : \mu_1(x) \neq \mu_2(x),$$

in which case $\mathbf{A} = (1, -1)$.

Suppose that $\{y_{ij}(x)\}$ take values in a function space $\mathcal{Y}$ spanned by a set of basis functions $\{\phi_k(x); k = 1, 2, \ldots\}$, e.g., Fourier basis, Wavelets, B-splines. We can then represent each $y_{ij}(x)$ by the linear combination $y_{ij}(x) = \sum_{k=1}^{\infty} c_{ijk}\phi_k(x)$, where $\{c_{ijk}; k = 1, \ldots, \}$ denote the sequence of basis coefficients. This allows us to transform the null hypothesis in (3.1) to the basis space. In particular, denote $\mu_i(x) = \sum_{k=1}^{\infty} u_{ik}\phi_k(x)$, $\mathbf{A}\,\boldsymbol{u}(x) = 0$ can be written as

$$A \begin{pmatrix} u_{11} & u_{12} & \cdots \\ u_{21} & u_{22} & \cdots \\ \vdots & \vdots & \ddots \\ u_{g1} & u_{g2} & \cdots \end{pmatrix} \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \end{pmatrix} = 0.$$

Testing the above null hypothesis is equivalent to testing the following sequence of hypotheses:

$$H_{0k} : \mathbf{A}\,\mathbf{u_k} = 0$$
$$H_{ak} : \mathbf{A}\,\mathbf{u_k} \neq 0,$$

for $k = 1, 2, \ldots$ Performing functional data testing in basis space brings several potential advantages: (1) **Dimension Reduction** Many basis provide sparse representations (aka,

sparse coding), which enables us to represent functional data with only a few nonzero coefficients, and thus substantially reduce the computation burden in functional testing. (2) **Decorrelation** Correlations between basis coefficients are often substantially reduced. Hence, basis representation reduces the difficulty of counting for correlation between test statistics. For example, many basis transformations (e.g., wavelets transform) make it possible to assume independence across basis coefficients, so that multiple testing adjustments that rely on independence assumption (e.g. step-down procedure based on Šidák method) may be used. (3) When compactly supported basis is used, significant components in basis-space testing can be mapped back to data domain for detection of significant regions.

3.3 Testing Procedure

As a demonstration example, we provide our testing procedure assuming by that the number of basis we finally use is K, with $K < \infty$.

Step 1. **Basis representation** Obtain the basis representation of each observation $\{y_{ij}(x);\ i = 1, \ldots, g;\ j = 1, \ldots, n_i\}$ using a common basis set $\{\phi_k(x)\}_{k\in\{1,\ldots,K\}}$. Denote the coefficients $\{c_{ijk}; i = 1, \ldots, g;\ j = 1, \ldots, n_i; k = 1, \ldots, K\}$. The coefficients can be obtained by taking the inner-product (for orthonormal basis) or minimizing an objective function (for basis such as B-splines).

Step 2. **Component-wise tests** Test $H_{0k} : \mathbf{A}\,\mathbf{u_k} = \mathbf{0}$ v.s. $H_{ak} : \mathbf{A}\,\mathbf{u_k} \neq \mathbf{0}$ using all coefficients $\{c_{ijk}\}$ for each fixed k. Denote the corresponding p-values $\{p_k;\ k = 1, ..., K\}$.

Step 3. **Sort p-values** Sort the p-values in increasing order: $p_{(1)} \leq p_{(2)} \leq ...p_{(K)}$, and let $\{r_1, r_2, ..., r_K\}$ denote the original index of $\{p_{(1)}, p_{(2)}, \ldots, p_{(K)}\}$ in the unsorted sequences $\{p_k, k = 1, \ldots, K\}$, so that $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_K}$.

Step 4. **P-value guided compression** For a pre-specified level τ, denote r_T by the index of largest p-value smaller than τ. We accept all tests with index $r_T, r_{T+1}, ..., r_K$, and keep the T tests with p-values smaller than τ, so that there are only T tests retained with original indices $r_1, \ldots, r_T$

Step 5. **Multiple testing adjustment** Perform multiple testing adjustment to the retained tests. For example, use Westfall-Young randomization adjustment to control FWER or use Benjamini-Hochberg test to control FDR, or etc. Denote the resulting adjusted p-values $\widetilde{p}_{(j)}, j = 1 \ldots T$ which corresponds to the tests indexed by $r_1, ...r_T$.

To make the final conclusion, we follow two equivalent processes. The first is a "step-down" procedure: if $\widetilde{p}_{(1)} \leq \alpha$, reject $H_{0,(1)}$ and continue (otherwise accept all null hypothesis and stop); if $\widetilde{p}_{(2)} \leq \alpha$, reject $H_{0,(2)}$ and continue (otherwise accept $H_{0,(j)}$ for $j >= 2$ and stop); and so on. The second is an "one-step" procedure: find the smallest j such that $\widetilde{p}_{(j)} > \alpha$; call it j^+; reject all $\{H_{0\,(j)}\}$ with $j < j^+$ and accept the rest.

3.4 Theoretical Results

In this section, we describe the main theoretical results through three theorems. The first two theorems consider the case where the Westfall-Young randomization method is applied to adjust p-values. In Theorem 3.1, we will show that the FWER is strongly controlled when we use countably many bases to represent the data, which suggests that the Westfall-Young randomization method is appropriate for functional data under the basis-space setup. In Theorem 3.2, we will show that the Westfall-Young randomization method achieves asymptotic optimality among all basis space tests that control FWER. In Theorem 3.3, we consider the general case of all types of adjustments and show that, under suitable regularity con-

ditions, appropriate compression in basis space leads to improved power as compared to point-wise testing in data domain or basis-space testing without compression.

3.4.1 FWER is strongly controlled with Westfall-Young randomization adjustment

We now consider applying Westfall-Young randomization adjustment as the multiple-correction method in Step 5 described in Section 3.3. The details of this specific testing procedures are available in Appendix A.1. Assume that there is a set of infinite dimensional bases. Consider family-wise error rate (FWE) control for the "step-down" approach. Let $S_0 = \{k : H_{0k} \text{ is true}, k \in \mathbb{Z}^+\}$ denote the index set of all true null hypotheses and assume that $S_0 \neq \varnothing$. We define FWE as the event $\{\text{Reject at least one } H_{0k}, k \in S_0 \mid H_0^{S_0}\}$, we claim that this event is equivalent to the event $\{\inf_{l \in S_0} P_l < \alpha \mid H_0^{S_0}\}$, where P_l denotes the random p-value for the l th hypothesis and $\alpha > 0$ is the pre-specified significant level. Therefore, to show that FWE is under control, we need to show that

$$\begin{aligned} &\Pr\left(\left\{\text{Reject at least one } H_{0k}, k \in S_0 \mid H_0^{S_0}\right\}\right) = \Pr\left(\left\{\inf_{l \in S_0} P_l < \alpha \mid H_0^{S_0}\right\}\right) \\ = \;& \Pr\left(\left\{\inf_{r_l \in S_0} \widetilde{P}_{(l)} < \alpha \mid H_0^{S_0}\right\}\right) \leq \alpha. \end{aligned}$$

Here, $\{r_1, r_2, \ldots\}$ are the original indices of the ordered p-values $\{p_{(1)}, p_{(2)}, \ldots\}$ and $\widetilde{P}_{(l)}$ denotes the random adjusted p-value for the ordered hypothesis $H_{0,(l)}$ (ordered according to the observed p-values). We show the equivalence of these two events by verifying two directions. First, if the event $\{\text{Reject at least one } H_{0k}, k \in S_0 \mid H_0^{S_0}\}$ holds, then $\exists\, j \in S_0$ such that $P_j < \alpha$, which implies that $\inf_{l \in S_0} P_l < \alpha$. On the other hand, if the event $\{\inf_{l \in S_0} P_l < \alpha \mid H_0^{S_0}\}$ holds, then there are two cases: (1) $\inf_{l \in S_0} P_l = P_j$ for some $j \in S_0$, in which case we will reject at least one hypothesis in S_0. (2) $\inf_{l \in S_0} P_l < P_j$ for all $j \in S_0$.

In this case, we either have $P_j < \alpha$ for some j, in which case we will reject at least one hypothesis in S_0; or $P_j \geq \alpha$ for all j, in which case $\inf_{j \in S_0} P_j \geq \alpha$, which contradict with the fact that $\inf_{l \in S_0} P_l < \alpha$.

Assumption 1 (Subset Pivotality for An Infinite Set of Hypotheses) *Denote the set of true null hypotheses by $S_0 = \{k : H_{0k}$ is true; $k \in \mathbb{Z}^+\}$ and assume that $S_0 \neq \varnothing$. Denote S any subset of the true null hypothesis, with $S \subseteq S_0$, the distribution of random p-values $\{P_k, k \in S\}$ is identical under the restriction $\cap_{i \in S} H_{0i}$ and $H_0^C = \cap_{i=1}^{\infty} H_{0i}$. In words, the distribution of any subvector of p-values is unaffected by the truth or falsehood of hypotheses corresponding to p-values not included in the subvector.*

Theorem 3.1. *Assume that the basis set $\{\phi_k\}$ for the function space $\mathcal{Y}$ contains infinite number of basis functions. Denote the set of true null hypotheses by $S_0 = \{k : H_{0k}$ is true; $k \in \mathbb{Z}^+\}$ and assume that $S_0 \neq \varnothing$. Under the subset pivotality assumption for an infinite set of hypotheses, the step-down adjusted p-value procedure controls the FWE in the strong sense, i.e.,*

$$Pr\left(\text{Reject at least one } H_{0k},\, k \in S_0 \mid H_0^{S_0}\right) \leq \alpha,$$

where α is a pre-specified error bound (significance level).

Proof. Let $p_0 = \inf_{l \in S_0} p_l$ denote the lower bound of all p-values corresponding to the true null hypotheses. Denote $J(p_0) = \{j \in \mathbb{Z}^+ : p_j \geq p_0\}$. The adjusted p-value for p_0 is defined to be

$$\widetilde{p}_0 = \max\left\{\left\{\widetilde{p}_{(l)} : p_{(l)} \leq p_0\right\}, \Pr\left(\inf_{l \in J(p_0)} P_l \leq p_0 \mid H_0^C\right)\right\}.$$

Note that the max operation is to guarantee the adjusted p-values follow the same the monotonicity as the original observed p-values. We use $\widetilde{P}_0$ to denote the random adjusted p-value (based on e.g., randomization).

First, we claim that the event $\{\widetilde{p}_0 < \alpha \mid H_0^C\}$ is equivalent to the event $\{p_0 < x^{\alpha}_{J(p_0)} \mid H_0^C\}$, where $x^{\alpha}_{J(p_0)}$ is the α quantile of $\{\inf_{l\in J(p_0)} P_l \mid H_0^C\}$. This can be easily shown in two directions: First, if $\{\widetilde{p}_0 < \alpha \mid H_0^C\}$ holds, then $\Pr\left(\inf_{l\in J(p_0)} P_l \le p_0 \mid H_0^C\right) < \alpha$ by the definition of $\widetilde{p}_0$, which implies that $p_0 < x^{\alpha}_{J(p_0)}$. On the other hand, if the event $\{p_0 < x^{\alpha}_{J(p_0)} \mid H_0^C\}$ holds, then $\Pr\left(\inf_{l\in J(p_0)} P_l \le p_0 \mid H_0^C\right) < \alpha$ by the definition of $x^{\alpha}_{J(p_0)}$. Furthermore, we also see that all elements of $\left\{\widetilde{p}_{(l)} : p_{(l)} \le p_0\right\}$ are less than α because of the design of the step-down procedure (otherwise, the algorithm has stopped before rejecting any hypothesis in $J(p_0)$). Therefore, by definition of $\widetilde{p}_0$, we see $\widetilde{p}_0 < \alpha$.

Furthermore, we see that $S_0 \subset J(p_0)$ because of the definition of p_0 and $J(p_0)$, this implies that $\inf_{l\in J(p_0)} P_l \le \inf_{l\in S_0} P_l$ hence the α quantiles satisfy $x^{\alpha}_{J(p_0)} \le x^{\alpha}_{S_0}$. Thus we have $\Pr(\inf_{l\in S_0} P_l < x^{\alpha}_{J(p_0)} \mid H_0^C) \le \Pr(\inf_{l\in S_0} P_l < x^{\alpha}_{S_0} \mid H_0^C)$.

Summarizing the above results, we see that

$$
\begin{aligned}
FWER &= \Pr\left(\text{Reject at least one } H_{0k},\, k \in S_0 \mid H_0^{S_0}\right) \\
&= \Pr\left(\widetilde{P}_0 < \alpha \mid H_0^{S_0}\right) \\
&= \Pr\left(\widetilde{P}_0 < \alpha \mid H_0^{C}\right) \qquad \text{(Subset Pivotality for Infinite Number of Hypotheses)} \\
&= \Pr\left(P_0 < x^{\alpha}_{J(p_0)} \mid H_0^C\right) \\
&= \Pr\left(\inf_{l\in S_0} P_l < x^{\alpha}_{J(p_0)} \mid H_0^C\right) \\
&\le \Pr(\inf_{l\in S_0} P_l < x^{\alpha}_{S_0} \mid H_0^C) \\
&= \alpha.
\end{aligned}
$$

This shows that the FWE under partial true null hypothesis is bounded by α. □

3.4.2 Asymptotic optimality with Westfall-Young adjustment

In this section, we examine the power of the testing procedure using Westfall-Young randomization method as compared with other FWER controlled adjusting methods. Meinshausen et al. [42] have proved in their work that the Westfall-Young permutation procedure has the asymptotic optimality, i.e., one cannot improve the power further when the number of tests goes to infinite. In this section, we will tailor their proof and show the asymptotic property of the proposed basis-space testing procedure.

The basic idea to show the asymptotic optimality is to compare the single-step Westfall-Young method with the oracle single-step testing which has the largest possible threshold for adjusted p-values. We will show that the threshold found by Westfall-Young method converges to the oracle threshold when the number of tests $K \to \infty$.

We first define the one-step oracle procedure to be compared with. Let W be the data matrix, which has the first column recording group information and the rest K columns recording all the coefficients $\{c_{ijk}\}, k = 1, \ldots, K$ corresponding to K basis components. We denote the true null index set by $S_0 \subseteq \{1, \ldots, K\}$. Under the true null hypotheses, the joint distribution of all K coefficients is denoted by P_K. Let $p_k(W)$ be the p-value corresponding to the kth test based on the data W. All possible p-values constitute the set $S_n \subseteq [0, 1]$. When applying the single-step procedure, we will reject all $H_{0,j}$, if $p_j(W)$ is smaller than an oracle threshold $c_{K,n}(\alpha)$, which is the α-quantile of $min_{k \in S_0} p_k(W)$ under P_K, i.e.,

$$c_{K,n}(\alpha) = \max\{s \in S_n : P_K(\min_{k \in S_0} p_j(W) \leq s) \leq \alpha\}$$

Now, we look at the single-step Westfall-Young randomization method. It assumes that under the complete null hypotheses H_0^C, the distribution of data W is invariant to a certain group of permutations $\mathcal{G}$. That is, for every permutation $g \in \mathcal{G}$, W and gW have the same

distribution. The single-step Westfall-Young threshold is defined by

$$\begin{aligned}\hat{c}_{K,n} &= \max\{s \in S_n : \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \mathbb{1}\{\min_{k=1,\ldots,K} p_k(gW) \leq s\} \leq \alpha\} \\ &= \max\{s \in S_n : P^*(\min_{k=1,\ldots,K} p_k(W) \leq s) \leq \alpha\},\end{aligned}$$

where P^* is the permutation distribution defined by $P^*(f(W) \leq x) = \frac{1}{\mathcal{G}}\sum_{g\in\mathcal{G}} \mathbb{1}\{f(gW) \leq x\}$ and $f(\cdot)$ is a function mapping the data W to the real line.

In order to show the asymptotic optimality of Westfall-Young method under the basis-space testing setup, we need to make some assumptions first. The assumptions are generally divided into two groups. Group A is related to the property of data and Group B is related to properties of the testing procedure

Assumption 2A.1 (Block-independence) There exists a partition $A_1, \ldots, A_{B_K}$ such that for every pair of permutations $\min_{\tilde{\theta}\in\{g,g'\}} \min_{k\in A_b\cap S_0} p_k(\tilde{g}W)$ are mutually independent, where B_K denotes the number of blocks. Without loss of generality, we assume that there is a at least one true null hypothesis in each block. This assumption means that even some tests may be correlated, but none of them is correlated with all tests. It is obviously true when orthonormal basis is applied. For the basis not completely orthonormal, for example, wavelet basis, it is still reasonable to assume it holds.

2A.2 (Sparsity) The number of true alternatives is relatively smaller than the number of blocks, i.e., $|S_0^C|/B_K \to 0$ as $K \to \infty$. This holds for most cases when we are trying to discover only a few regions among dense measurements.

2A.3 (Block-size) The maximum of a block size is of smaller order than the square root of the number of blocks. If we use the orthonormal basis, the maximum size of a block will be 1. Then we have $1/K \to 0$ as $K \to \infty$, which matches this assumption. When applying

compactly supported basis, we are also able to assume that the range of independence is not too large.

2B.1 (Permutation pivotality) Under the true null hypothesis, the distribution of p-values will not change no matter how we permute the data. Let G be a random permutation picked from $\mathcal{G}$. Under $H_0^{S_0}$, $\{p_k(W), k \in S_0\}$ and $\{p_k(GW), k \in S_0\}$ have the same joint distribution.

2B.2 (Bounded permutation distribution) There exists a constant $r < \infty$, s.t., for $s = c_{K,n}(\alpha) \in S_n$ and all W,

$$\frac{s}{r} \leq P^*(p_k(W) \leq s) \leq rs \ \forall k = 1, \ldots, K.$$

2B.3 (Uniformly distributed p-values under true null) The p-values corresponding to true null hypotheses are uniformly distributed. That is,$\forall k \in S_0$ and $\forall s \in S_n P_K(p_k(W) \leq s) = s$

Based on Assumption 2, we could claim in Theorem 3.2 that the Westfall-Young threshold $\hat{c}_{K,n}$ will converge to the oracle threshold $c_{K,n}$ when K goes to infinity. We prove it generally following the proof given by Meinshausen et al. [42] and the details are given in Appendix A.2.

Theorem 3.2. *For any* $\alpha \in (0, 1)$ *and any* $\delta \in (0, \alpha)$

$$P_K\{\hat{c}_{K,n}(\alpha) \geq c_{K,n}(\alpha - \delta)\} \to 1 \ \textit{as} \ K \to \infty$$

Note that the sample size n can be fixed and does not need to tend to ∞. However, if the range S_n is discrete, n must increase with K to avoid a trivial results where $c_{K,n}(\alpha - \delta)$ vanishes.

3.4.3 Appropriate compression in basis space leads to improved empirical power

In general, when we use parsimonious basis to represent data, it is reasonable for us to assume that only a part of the components carries useful information. We could truncate the basis components that carry non-relevant information. Here we propose to truncate the tests based on the unadjusted p-values with a pre-specified level of significance. The truncated components are expected to contain much less true alternatives than the full set of tests. With this assumption, in **Theorem 3.3**, we will show that the testing procedure based on the compressed sets of coefficients can be more powerful than the one without compression under certain regularity conditions. To show this, we first introduce some notations. Suppose that we truncate the tests at T and recall that the kth largest p-value is denoted by $p_{(k)}$, we denote the adjusted p-value of $p_{(k)}$ based on the truncated set of tests by $\tilde{p}^t_{(k)}$:

$$\tilde{p}^t_{(k)} = P_K(\min_{1\le j\le T} p_{(j)} \le p_{(k)}).$$

Recall that based on the full set of tests, the adjusted p-value of $p_{(k)}$ is defined as:

$$\tilde{p}_{(k)} = P_K(\min_{j\in\{1,\dots,K\}} p_{(j)} \le p_{(k)}).$$

Let S_a be the set containing the ranks of unadjusted p-values of the tests in the true alternative set S_0^C. [1]

Theorem 3.3. *The empirical power of truncated tests is greater than the empirical power*

[1]For example, suppose we have three tests in total, $S_0 = \{1\}$, and $S_0^C = \{2, 3\}$. The ranks of the three unadjusted p-values are 2, 1, and 3. Then, $S_a = \{1, 3\}$

of the full tests. That is:

$$\frac{\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\tilde{p}^t_{(k)} < \alpha\}/ \mid S_a \cap \{1,...,T\} \mid}{\sum_{k\in S_a} \mathbb{1}\{\tilde{p}_{(k)} < \alpha\}/ \mid S_a \mid} > 1.$$

Proof.

$$\begin{aligned}
&\frac{\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\tilde{p}^t_{(k)}<\alpha\}/|S_a\cap\{1,...,T\}|}{\sum_{k\in S_a} \mathbb{1}\{\tilde{p}_{(k)}<\alpha\}/|S_a|} \\
&= \frac{\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\tilde{p}^t_{(k)}<\alpha\}}{\sum_{k\in S_a} \mathbb{1}\{\tilde{p}_{(k)}<\alpha\}} \cdot \frac{|S_a|}{|S_a\cap\{1,...,T\}|} \\
&= \frac{\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\tilde{p}^t_{(k)}<\alpha\}}{\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\tilde{p}_{(k)}<\alpha\}+\sum_{k\in S_a\cap\{T+1,...\}} \mathbb{1}\{\tilde{p}_{(k)}<\alpha\}} \cdot \frac{|S_a\cap\{1,...,T\}|+|S_a\cap\{T+1,...\}|}{|S_a\cap\{1,...,T\}|}.
\end{aligned}$$

Let:

$$\begin{aligned}
a &= \textstyle\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\widetilde{p_{(k)}} < \alpha\}; \\
a_t &= \textstyle\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\widetilde{p^t_{(k)}} < \alpha\}; \\
x &= \textstyle\sum_{k\in S_a\cap\{T+1,...\}} \mathbb{1}\{\widetilde{p_{(k)}} < \alpha\}; \\
a' &= \textstyle\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\widetilde{p_{(k)}} \geq \alpha\}; \\
a'_t &= \textstyle\sum_{k\in S_a\cap\{1,...,T\}} \mathbb{1}\{\widetilde{p^t_{(k)}} \geq \alpha\}; \\
x' &= \textstyle\sum_{k\in S_a\cap\{T+1,...\}} \mathbb{1}\{\widetilde{p_{(k)}} \geq \alpha\}.
\end{aligned}$$

Notice that, as $\tilde{p}^t_{(k)}$ stands for the adjusted p-value of the T retained tests, it is not defined

for the rest of the tests. The ratio above becomes:

$$\begin{aligned}\frac{a_t}{a+x}\cdot\frac{(a+a')+(x+x')}{a_t+a_t'} &= \frac{\frac{(a+x)+(a'+x')}{a+x}}{\frac{a_t+a_t'}{a_t}} \\ &= \frac{1+\frac{a'+x'}{a+x}}{1+\frac{a_t'}{a_t}} \\ &= 1+\frac{\frac{a'+x'}{a+x}-\frac{a_t'}{a_t}}{1+\frac{a_t'}{a_t}}\end{aligned}$$

Thus, we only need to show that

$$\frac{a'+x'}{a+x}-\frac{a_t'}{a_t}>0.$$

We have,

$$\frac{a'+x'}{a+x}-\frac{a_t'}{a_t}=\frac{a_ta'+a_tx'-aa_t'-a'x}{a_t(a+x)}$$

If $a_ta'+a_tx'-aa_t'-a'x>0$, then $\frac{a'+x'}{a+x}-\frac{a_t'}{a_t}>0$. This can be interpreted as: the proportion of accepted tests based on all tests is larger than that of the retained tests, which holds if we assume that the probability of rejection is much smaller in the truncated part than that in the selected part. As compression usually achieves dimension reduction by throwing out high frequency (noisy) components, this assumption is intuitive and easy to fulfill. □

3.5 Simulation Studies

We perform two simulation studies, one for 1-dimensional case and one for 3-dimensional case to evaluate the performance of the proposed testing procedure. In both cases we adopt the two-sample t-test to calculate unadjusted p-values for testing group-mean differences. In each case, one group has zero mean and the other group has a certain pattern. We apply wavelet basis representation using daubechies wavelets, which is a multi-resolution and com-

pactly supported basis. Wavelet representation allows us to map the significant components obtained in basis space back in data domain for the purpose of detecting significant regions. Since our ultimate goal is making decision and detecting significant region in data domain, we first define two sets of summary statistics for comparison purposes, one in data domain for evaluation of the region detection and the other in basis domain for the evaluation of the theoretical properties in Section 3.5.1. The summary statistics are False Discovery Rate (FDR), power, specificity and FWER.

In order to compare the performance of our method with other tests in both data and basis domain, we mainly consider the following six specific testing procedures:

1 Basis-space testing with Westfall-Young randomization method, controlling FWER in basis domain at 0.05;

2 Basis-space testing with Holm's correction, controlling FWER in basis domain at 0.05;

3 Basis-space testing with BH correction, controlling FDR in basis domain at 0.05;

4 Point-wise testing with Westfall-Young randomization method, controlling FWER in data domain at 0.05;

5 Point-wise testing with Holm's correction, controlling FWER in data domain at 0.05;

6 Point-wise testing with BH correction, controlling FDR in datadomain at 0.05.

Section 3.5.2 demonstrates the results of the one-dimensional simulation study, in which the design group has a mean curve consisting of a lower sine wave followed by a spike. We compare it with other tests of interest by using the summary statistics defined in Section 3.5.1. We also visualize how the proposed method works in detecting region of significance. Section 3.5.3 provides the evaluations for 3-dimensional case. In the 3-dimensional simulation

study, our simulated example mimics the neuroimaging data by defining different pattern of abnormality in the brain, and the data is characterized by a $220 \times 220 \times 220$ cube. The design group has a mean that consists of three different types of abnormal variations located in different areas of a brain.

3.5.1 An overview of simulation study

Since our test is performed in basis space but regions are detected and interpreted in the data domain, we will evaluate our methods in both data domain and in basis domain. In order to do this, in this section we define FDR, power, specificity and FWER in both data domain and basis domain as follows.

Summary Statistics in Basis Domain The above listed summary statistics are easy to define in basis domain. The only thing to mention is that, for testing in basis domain, true differences that are very small should be treated as zero. Therefore, we need to set a threshold so that components with differences less than the threshold are treated as true null hypotheses. In data domain, since the data is continuous, we also need a threshold to define on which grid points that alternative hypothesis holds. For example, in the EEM data example(Section 1.2.3), we usually consider the area with difference larger than 0.02 as the region of true alternatives. In the basis domain, we use a value close to zero as a threshold, denoted as ξ. Based on ξ, first we calculate the true mean difference $\boldsymbol{\mu}$ in data domain and get the coefficient vector of its basis representation, denoted as $\boldsymbol{c} = \{c_1, \ldots, c_K\}$. Thus $S_{0,\xi} = \{k, |c_k| < \xi\}$ is the set of true null and its complimentary part $S_{0,\xi}^C = \{k, |c_k| \geq \xi\}$ is the true alternative set. Denote the rejection set by R and acceptance set by R^c. We define the summary statistics in

basis domain as follows:

$$FDR_\xi = \frac{|R \cap S_{0,\xi}|}{|R|}$$
$$Power_\xi = \frac{|R \cap S_{0,\xi}^C|}{|S_0^C|}$$
$$Specificity_\xi = \frac{R^C \cap S_{0,\xi}}{|S_{0,\xi}|}$$
$$FWER_\xi = Pr(|R \cap S_{0,\xi}| \geq 1)$$

Summary Statistics in Data Domain In data domain, we denote the threshold of defining true difference by ζ. ζ is a fixed number. For example, in our 3-D simulation, the minimum of the added pattern is 0.6 for the design group. Thus, $\zeta = 0.6$. In general, it could also be a self-specified number, as sometimes the pattern of contrast is a continuous function such as our 1-d simulation example. In the 1-d case, we always set ζ to be a reasonably small value. Suppose our data is obtained on N grid points. We first calculate the true mean difference as $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_N\}$. Similarly, we denote the index set of null hypotheses by $S_{0,\zeta} = \{i, |\mu_i| < \zeta\}$ and its complimentary part by $S_{0,\zeta}^C = \{i, |\mu_i| \geq \zeta\}$. Now we need to determine which hypotheses in data domain will be flagged as "significant" based on the test results in data domain. To do this, we first obtain the coefficient vector $\boldsymbol{c} = \{c_1, ..., c_K\}$ for $\boldsymbol{\mu}$. We retain significant coefficients while setting the non-significant ones to zero, which gives us a new coefficient vector $\boldsymbol{c'} = \{c'_1, ...c'_K\}$. Specifically,

$$c'_k = \begin{cases} c_k, & if\ H_{0,k}\ is\ rejected; \\ 0, & otherwise \end{cases};$$

We then map back $\boldsymbol{c'}$ to data domain and get the reconstructed mean difference $\boldsymbol{\mu'}$. The next step is to set a threshold ζ' for flagging differences using $\boldsymbol{\mu'}$. Usually, we

set $\zeta' = \zeta$. However, if we assume magnitude of signal may change substantially after mapping back to data domain, ζ' could also be set to a reasonably small value. Next, we define the region of significance based on the testing result by $F_{\zeta'} = \{i, |\mu'_i| \geq \zeta'\}$; the index set for unflagged region is defined by $F^C_{\zeta'} = \{i, |\mu'_i| < \zeta'\}$. The summary statistics in data domain for basis-space testing method are then defined by:

$$
\begin{aligned}
FDR_{\zeta,\zeta'} &= \frac{|F_{\zeta'} \cap S_{0,\zeta}|}{|F_{\zeta'}|}, \\
Power_{\zeta,\zeta'} &= \frac{|F_{\zeta'} \cap S^C_{0,\zeta}|}{|S^C_0|}, \\
Specificity_{\zeta,\zeta'} &= \frac{F^C_{\zeta'} \cap S_{0,\zeta}}{|S_{0,\zeta}|}, \\
FWER_{\zeta,\zeta'} &= Pr(|F_{\zeta'} \cap S_{0,\zeta}| \geq 1).
\end{aligned}
$$

In the above formulae, if we use the same value for ζ and ζ', the footnote of the statistics could be simplified by using ζ alone, e.g., $FDR_{\zeta,\zeta'} = FDR_\zeta$ if $\zeta' = \zeta$.

In the next two sections, we report our simulation results using the above defined summary statistics. Here, we only defined statistics for the basis-space tests. For the point-wise tests, the summary statistics in basis domain are not meaningful. Thus, we won't calculate them.

3.5.2 Simulation study for one-dimensional data

We first present a simulation study with generated 1-d curves from two groups. The control group has zero mean $\mu_0(x) = 0$ and the design group has a mean function $\mu_1(x)$ defined by:

$$0.1 \times sin(4\pi(x-1)) + 0.025 \times \beta(x - 0.35, 1000, 1000),$$

which is plotted in Figure 3.2. As demonstrated in Figure 3.2, this mean function consists of a sine wave followed by a beta spike.

For the control group, we first generate zero-mean Gaussian Processes evaluated on 1400 equally spaced grid points on [-0.2,1.2] with zero mean and covariance function $Cov(x_1, x_2) = 0.04exp(-\frac{(x_1-x_2)^2}{0.95^2})$. We then cut 200 points on each end to obtain 1000 equally spaced values on [0,1]. This approach helps to generate smooth curves. Each group has 250 observations. We adopt wavelet representation for simulated data using Daubechies wavelets with the maximal number of vanishing moments being 4 (i.e., db4). The number of resolution levels is set to be $J = 3$, and the boundary extension mode is set to be periodic. At every wavelet component, we calculate the unadjusted p-values based on two sample t-test. The tests with unadjusted p-values greater than the threshold τ (which is set to be 0.05 and 0.01) are truncated. Finally we obtain the adjusted p-values with Westfall-Young randomization method, Holm's method and BH correction. The resulted adjusted p-values in basis domain using these three methods are shown in Figure 3.1. In Figure 3.1, we mark these three methods with different colors and mark the significant level 0.05 with a red dashed line. From Figure 3.1, we observed the following patterns. (1) All three methods have similar performance in components of high resolution and (2) in components of low resolution, the original pattern is still preserved, the FDR control method wavelet-BH results in more adjusted p-values that are lower than α. However, for the two FWER control methods, using Westfall-Young randomization adjustment identifies more true significant components than using the Holm's method.

To further compare our basis-space testing procedure with the data domain point-wise tests, in data domain, we first apply point-wise two sample t-test to obtain the unadjusted p-values. We then adjust these p-values by using the same three correction methods similar to basis-space testing. Last, set ζ to be $1e-6$ for extremely small diefferential effect and 0.01 for moderate differential effect respectively. We set $\zeta' = \zeta$ in data domain. The summary statistics of theses six procedures of interest from 1000 simulations are summarized in Table

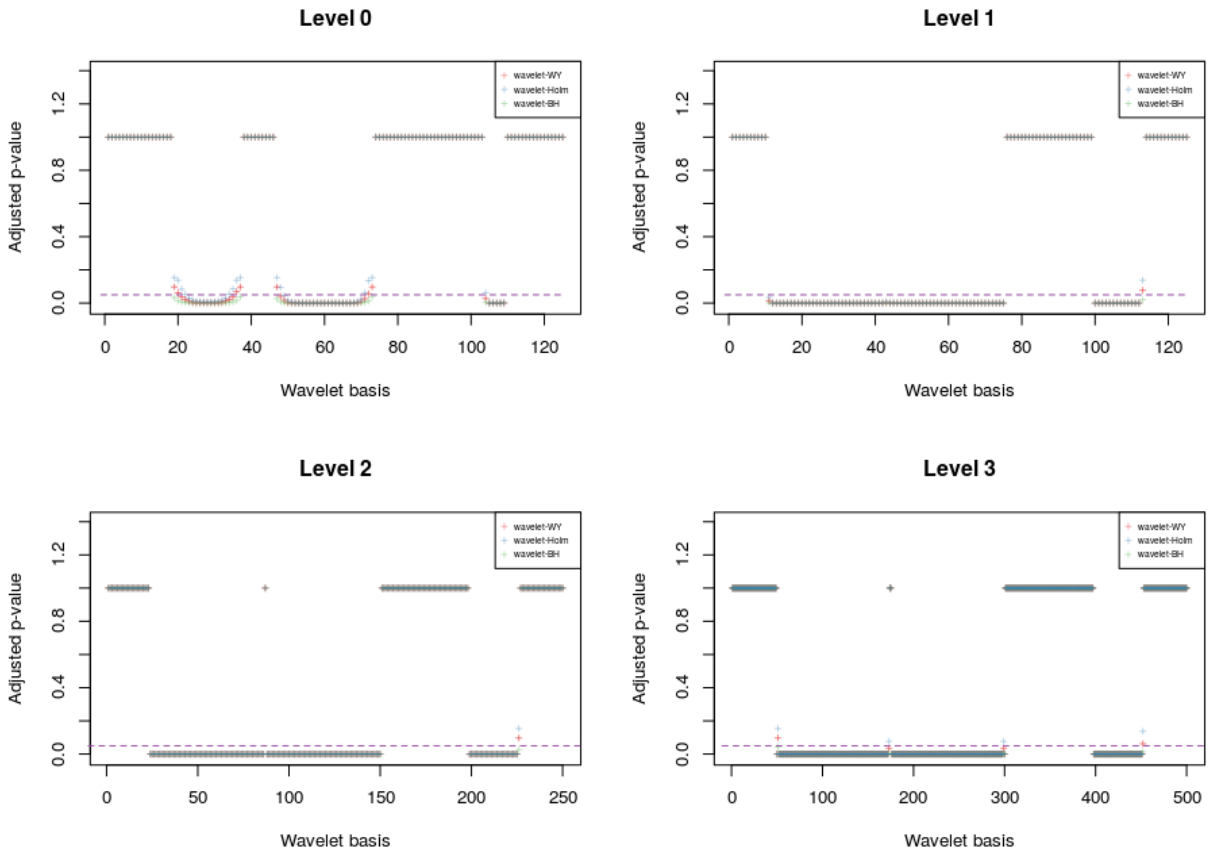

Figure 3.1: Plot of adjusted p-values in basis space at each wavelet component: x-axis indicates the relative locations of the wavelet components. Level-0 consists of coarse scale approximation coefficients. Level 1-3 consist of detailed coefficients at three resolution scales. Red crosses represent p-value based on Westfall-Young randomization method; Blue crosses represent p-value based on Holm's method; Green crosses represent p-value based on BH correction.

3.1. From Table 3.1 we can see that:

- The overall performance of basis-space testing is better than the point-wise testing, with respect to power, which is consistent with Theorem 3.2.

- The basis-space testing procedure with Westfall-Young randomization adjustment could control FWER around the given significance level $\alpha = 0.05$ in the basis domain, espe-

cially when heavier compression is performed.

- For moderate differences present, the basis-space testing could detect difference with more power than point-wise test.

- For the purpose of detecting small differences, the basis-space testing controls FDR well while may require larger sample size or higher signal-to-noise ratio to control the family-wise error.

- Among the three basis-space testing procedures examined, the power of Westfall-Young randomization procedure, is very close to the FDR controlled correction and more powerful than the Holm's method. The possible reason is that, although correlations between wavelet coefficients are substantially reduced, there still exists correlations between some wavelet components[66]. Thus, there are still some correlated tests, and Holm's method would lose power when correlations exist.

Finally, we will investigate the performance of these six testing procedures in detecting regions of difference based on one simulation. The result is shown in Figure 3.2. We mark the true region of difference with $|\mu_1(x)-\mu_0(x)| > \zeta$ using black crosses along the curves. For different levels of ζ, we mark the region detected by each method using different colors under the curve of mean difference. From Figure 3.2, we see that, in general, the three basis-space based methods can detect more regions than the point-wise tests. Among the point-wise tests, Westfall-Young randomization test and BH method have similar performance while Holm's method misses one part of the negative sine wave. For moderate difference, we can see that, using Westfall-Young randomization adjustment gives almost the same region as the FDR control method. Both detected regions that are wider than the region detected by Holm's method.

			Data Domain				Wavelet Domain				
	Method	τ	FDR	power	specificity	FWER	FDR	power	specificity	FWER	
$\zeta = 1e-6$	WY	-	0.009 (0.051)	0.796 (0.014)	0.982 (0.105)	0.037	-	-	-	-	$\xi = 1e-11$
	Holm's		0.000 (0.000)	0.705 (0.009)	1.000 (0.000)	0	-	-	-	-	
	BH		0.0144 (0.068)	0.799 (0.018)	0.969 (0.148)	0.047	-	-	-	-	
	wWY	0.05	0.062 (0.006)	0.999 (0.001)	0.893 (0.011)	-	0.007 (0.031)	0.917 (0.006)	0.984 (0.071)	0.094	
	wHolm's		0.063 (0.002)	0.999 (0.001)	0.891 (0.004)	-	0.000 (0.002)	0.913 (0.004)	0.999 (0.004)	0.119	
	wBH		0.061 (0.002)	0.999 (0.001)	0.896 (0.017)	-	0.020 (0.052)	0.922 (0.001)	0.955 (0.121)	0.289	
	wWY	0.01	0.063 (0.004)	0.999 (0.001)	0.891 (0.007)	-	0.004 (0.022)	0.914 (0.005)	0.992 (0.050)	0.050	
	wHolm's		0.063 (0.000)	0.999 (0.001)	0.891 (0.005)	-	0.000 (0.003)	0.913 (0.003)	0.999 (0.006)	0.050	
	wBH		0.064 (0.003)	0.999 (0.001)	0.891 (0.004)	-	0.004 (0.022)	0.914 (0.005)	0.992 (0.050)	0.050	
$\zeta = 0.01$	WY	-	0.011 (0.060)	0.888 (0.010)	0.981 (0.106)	0.103	-	-	-	-	$\xi = 1e-8$
	Holm's		0.000 (0.000)	0.789 (0.010)	1.000 (0.000)	0.001	-	-	-	-	
	BH		0.017 (0.079)	0.891 (0.009)	0.969 (0.147)	0.117	-	-	-	-	
	wWY	0.05	0 (0)	0.914 (0.014)	1.000 (0.000)	0.002	0.008 (0.033)	0.951 (0.005)	0.984 (0.000)	0.103	
	wHolm's		0 (0)	0.899 (0.025)	1.000 (0.000)	0.001	0.000 (0.002)	0.948 (0.004)	0.999 (0.071)	0.144	
	wBH		0 (0)	0.893 (0.024)	1(0)	0	0.022 (0.056)	0.955 (0.006)	0.954 (0.122)	0.222	
	wWY	0.01	0.000 (0.000)	0.897 (0.010)	1.000 (0.000)	0	0.004 (0.024)	0.947 (0.004)	0.992 (0.050)	0.056	
	wHolm's		0.000 (0.000)	0.896 (0.012)	1.000 (0.000)	0	0.001 (0.003)	0.947 (0.003)	0.999 (0.006)	0.056	
	wBH		0.000 (0.000)	0.897 (0.010)	1.000 (0.000)	0	0.004 (0.024)	0.947 (0.004)	0.992 (0.050)	0.056	

Table 3.1: Summary statistics in 1-D simulation study: Left table: statistics calculated in data domain; right table: statistics calculated in wavelet domain. The statistics include mean (numbers before parentheses) and standard deviation (numbers in the parentheses) of the FDR, power, and specificity. Here, they are calculated following the definitions in Section 3.5.1 based on 1000 simulations. FWER is calculated as the proportion of at least one wrong rejection among all 1000 simulations. We consider different levels of compression for original p-value:τ, different levels of differences in data domain ζ and different levels of true difference in wavelet domain:ξ for all procedures of interests. WY, Holm's, BH, wWY, wHolm's, wBH stand for point-wise tests with West-fall Young randomization adjustment, Holm's correction, BH method and wavelet-based tests with Westfall-Young randomization adjustment, Holm's correction and BH method respectively.

3.5.3 A simulation study for three-dimensional data

We perform 3-D simulation study by generating data that mimick the MRI image. The relationship between the MRI image and disease status is modeled by:

$$y_i(d) = \mu(d) + X_i\alpha(d) + \tau_i(d) + e_i(d),$$

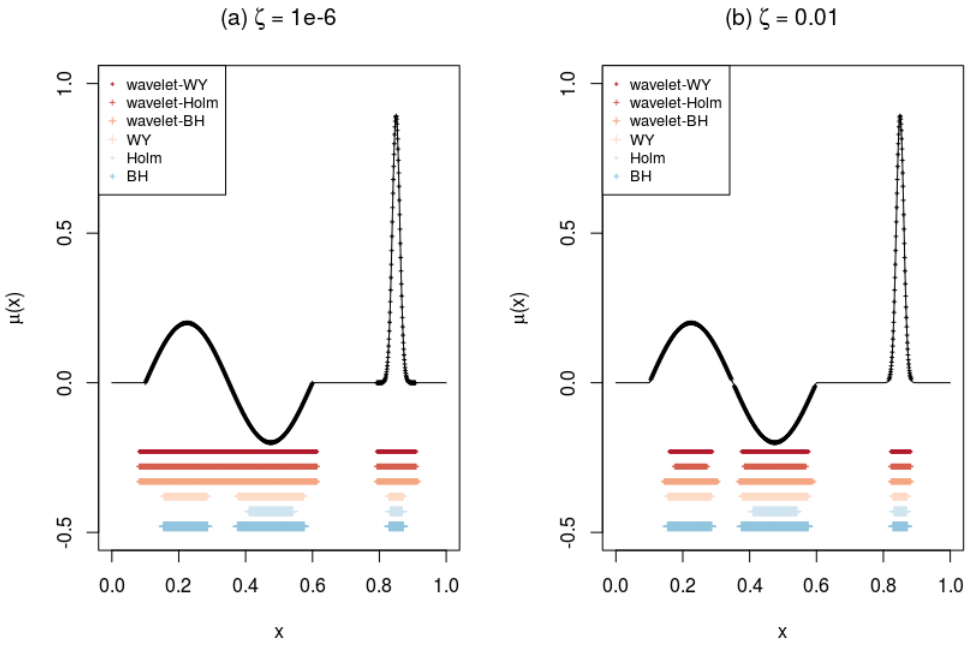

Figure 3.2: Detected regions of difference by applying different testing procedures in the 1-d simulation. We mark the region detected by different method with different colors below the curve of mean difference. The true region of difference according to threshold ζ is marked by black cross on the curve. (a) shows results for detecting small true divergence with $\zeta =$ 1e-6. (b) shows results for detecting lager true divergence with $\zeta = 0.01$.

where $\tau_i(d)$ is the individual random effect, $\alpha(d)$ denotes the disease-related effect, $\mu(d)$ is the population mean, and e_i are i.i.d. white noise. Here we consider the test of mean differences, which corresponds to testing whether $\alpha(d)$ equals zeros, where X_i is a binary variable indicating the disease status.

We set the group mean $\mu(d)$ for the control group be a reference image obtained from the real data. This reference image is the sample average of all images in the NL group, measured on a $220 \times 220 \times 220$ equally gridded cube. Based on $\mu(d)$, we then generate three additive patterns for the mean of the design group in 3 separate regions of the brain, denoted by $B_1(d)$, $B_2(d)$ and $B_3(d)$, which reflect different types of abnormality related to the disease. Therefore, the mean for the design group is $\mu_1(d) = \mu(d) + B_1(d) + B_2(d) + B_3(d)$. The

patterns of $B_1(d)$, $B_2(d)$ and $B_3(d)$ are displayed in the first row of Figure 3.3. Specifically, $B_1(d)$ demonstrates an area with linearly increasing value along one dimension, $B_2(d)$ reflects a ball shape with the largest value in the middle and $B_3(d)$ reflects a star shape respectively. The minimum non-zero value of these patterns is 0.6.

The random effect for the ith individual i is generated by $\tau_i(d) = \xi_{i1}\phi_1(d)+\xi_{i2}\phi_2(d)+\xi_{i3}\phi_3(d)$, where $\{\phi_1, \phi_2, \phi_3\}$ is the set of eigen-basis and $\xi_{i1}, \xi_{i2}, \xi_{i3}$ are normal random variables with mean zero and variances $\lambda_1, \lambda_2, \lambda_3$ respectively. The eigen-basis are specified by $\phi_1(d) = f_1(d_x)$, $\phi_2(d) = f_2(d_y)$ and $\phi_3(d) = f_3(d_z)$, where f_1, f_2, f_3 are i.i.d Gaussian Processes with mean 0 and covariance kernel $k(s,t) = \sigma^2 exp\{-||s-t||^2/2\rho^2\}$. We set $\sigma^2 = 1$ and $\rho^2 = 100$. The white noise term $e_i(d)$ follows i.i.d $N(0,1)$.

Following the above specifications, we generate 50 samples for each group. In this study, we only consider the basis space based tests because the dimension of the point-wise tests goes beyond 10^8, which requires 325 megabytes to save one sequence of p-value. If we perform 1000 permutations for the Westfall-Young randomization adjustment, we need a memory space larger than 325 gigabytes. Therefore, the point-wise tests cost too much computation resource which goes beyond what a personal pc can handle.

For the basis-space test, we adopt wavelet basis to represent data. We use Daubechies wavelets with maximal number of vanishing moments six (i.e., db6). The number of resolution levels is set to be $J = 4$. The boundary extension mode is set to be periodic. For each wavelet component, we use two sample t-test to calculate the unadjusted p-values. We consider three scenarios with the truncation threshold τ set to be 1e-4, 1e-5, and 1e-6 respectively. Tests with p-values larger than τ were truncated. Based on the retained tests, we obtain the adjusted p-values using Westfall-Young randomization method or Holm's method controlling FWER at 0.05 and BH controlling FDR at 0.05. The performances are evaluated through the same summary statistics as in the one-dimensional simulation. Results

are shown in Table 3.2. In this simulation, we set $\zeta = 0.6$ and obtain flagged regions with different ζ''s. From Table 3.2, we see that, the three adjustments have just the same performance in both data domain and time domain. If we closer at the results, we find that the unadjusted p-values are not the same. They all have great power in identifying region of interest.

In Figure 3.3 we pick some 2-D slices of brain to show the performance of basis-space testing in identifying the region of difference. Here, we set the threshold of compression to be $\tau = 1e - 4$. Based on one simulation, we see that, the proposed method is able to identify different shapes of patterns. In addition, the reconstructed contrast effect has similar value as the true difference in magnitude. This suggests that, it is reasonable to apply a threshold in data domain directly on the reconstructed mean to flag the region of difference.

3.6 Application

We apply the proposed basis-space testing to two real datasets reviewed in Section 1.2.3 and Section 1.2.4. For both datasets, we still use wavelet representation. Our main task here is to identify regions on biomedical signals/images that reflect differences between groups.

3.6.1 Fluorescence spectroscopy data for pre-cancer diagnosis

For the EEM data described in Section 1.2.3, we represent data at each level of emission wavelength separately and concatenate the resulting coefficients from all emission wavelength levels. At each wavelet component, two sample t-test is performed to obtain raw p-values. We set τ, the threshold used in p-value guided compression, to be 0.01. We then perform the Westfall-Young adjustment based on p-values from the retained wavelet components. In

			Data Domain			Wavelet Domain				
	Method	τ_p	FDR	power	specificity	FWER	FDR	power	specificity	FWER	
$\zeta' = 0.6$	wWY	1e-4	0 (0)	0.986 (0.001)	1 (0)	0	0.076 (0.003)	0.928 (0.001)	1.000 (<1e-4)	-	$\xi = 1e-4$
	wHolm's		0 (0)	0.986 (0.001)	1 (0)	0	0.076 (0.003)	0.928 (0.001)	1.000 (<1e-4)	-	
	wBH		0 (0)	0.986 (0.001)	1 (0)	0	0.076 (0.003)	0.928 (0.001)	1.000 (<1e-4)	-	
	wWY	1e-5	0(0)	0.986 (0.001)	1 (0)	0	0.070 (0.003)	0.918 (0.015)	1.000 (<1e-4)	-	
	wHolm's		0 (0)	0.986 (0.001)	1 (0)	0	0.070 (0.003)	0.918 (0.015)	1.000 (<1e-4)	-	
	wBH		0 (0)	0.986 (0.001)	1 (0)	0	0.070 (0.003)	0.918 (0.015)	1.000 (<1e-4)	-	
	wWY	1e-6	0 (0)	0.986 (0.001)	1 (0)	0	0.068 (0.003)	0.907 (0.015)	1.000 (<1e-4)	-	
	wHolm's		0 (0)	0.986 (0.001)	1 (0)	0	0.068 (0.003)	0.907 (0.015)	1.000 (<1e-4)	-	
	wBH		0 (0)	0.986 (0.001)	1 (0)	0	0.068 (0.003)	0.907 (0.015)	1.000 (<1e-4)	-	
$\zeta' = 1.2$	wWY	1e-4	0 (0)	0.862 (0.001)	1 (0)	0	0.028 (0.001)	0.807 (0.013)	1.000 (<1e-4)	-	$\xi = 1e-6$
	wHolm's		0 (0)	0.862 (0.001)	1 (0)	0	0.028 (0.001)	0.807 (0.013)	1.000 (<1e-4)	-	
	wBH		0 (0)	0.862 (0.001)	1 (0)	0	0.028 (0.001)	0.807 (0.013)	1.000 (<1e-4)	-	
	wWY	1e-5	0 (0)	0.862 (0.001)	1 (0)	0	0.023 (0.001)	0.797 (0.001)	1.000 (<1e-4)	-	
	wHolm's		0 (0)	0.862 (0.001)	1 (0)	0	0.023 (0.001)	0.797 (0.001)	1.000 (<1e-4)	-	
	wBH		0 (0)	0.862 (0.001)	1 (0)	0	0.023 (0.001)	0.797 (0.001)	1.000 (<1e-4)	-	
	wWY	1e-6	0 (0)	0.862 (0.001)	1 (0)	0	0.022 (0.001)	0.787 (0.014)	1.000 (<1e-4)	-	
	wHolm's		0 (0)	0.862 (0.001)	1 (0)	0	0.022 (0.001)	0.787 (0.014)	1.000 (<1e-4)	-	
	wBH		0 (0)	0.862 (0.001)	1 (0)	0	0.022 (0.001)	0.787 (0.014)	1.000 (<1e-4)	-	
$\zeta' = 1.8$	wWY	1e-4	0 (0)	0.718 (0.001)	1 (0)	0	0.005 (0.001)	0.698 (0.012)	1.000 (<1e-4)	-	$\xi = 0$
	wHolm's		0 (0)	0.718 (0.001)	1 (0)	0	0.005 (0.001)	0.698 (0.012)	1.000 (<1e-4)	-	
	wBH		0 (0)	0.718 (0.001)	1 (0)	0	0.005 (0.001)	0.698 (0.012)	1.000 (<1e-4)	-	
	wWY	1e-5	0 (0)	0.718 (0.001)	1 (0)	0	0.000 (0.000)	0.689 (0.012)	1.000 (<1e-4)	-	
	wHolm's		0 (0)	0.718 (0.001)	1 (0)	0	0.000 (0.000)	0.689 (0.012)	1.000 (<1e-4)	-	
	wBH		0 (0)	0.718 (0.001)	1 (0)	0	0.000 (0.000)	0.689 (0.012)	1.000 (<1e-4)	-	
	wWY	1e-6	0 (0)	0.718 (0.001)	1 (0)	0	0.000 (0.000)	0.680 (0.013)	1.000 (<1e-4)	0.3	
	wHolm's		0 (0)	0.718 (0.001)	1 (0)	0	0.000 (0.000)	0.680 (0.013)	1.000 (<1e-4)	0.3	
	wBH		0 (0)	0.718 (0.001)	1 (0)	0	0.000 (0.000)	0.680 (0.013)	1.000 (<1e-4)	0.3	

Table 3.2: Summary statistics from 3-D simulation study. The three wavelet-based testing of interest are Westfall-Young randomization test and Holm's method, both of which controls FWER at 0.05 in wavelet domain, and BH correction, which controls FDR at 0.05 in wavelet domain. Left table shows the statistics in data domain and right table shows the statistics in wavelet domain. The means of FDR, power and specificity are demonstrated, and the standard deviations are given inside the parentheses. FWER is calculated as the proportion of simulation with at least one wrong rejection. Result is obtained based on 500 simulations.

order to compare the performance of our test with the point-wise test in data domain, we apply point-wise Westfall-Young test in data domain as well. Note that as the intensity of light is always a positive number, the distribution of data right-skewed. This violates the distribution assumption for two sample t-test. Thus, for the point-wise test, we applied the

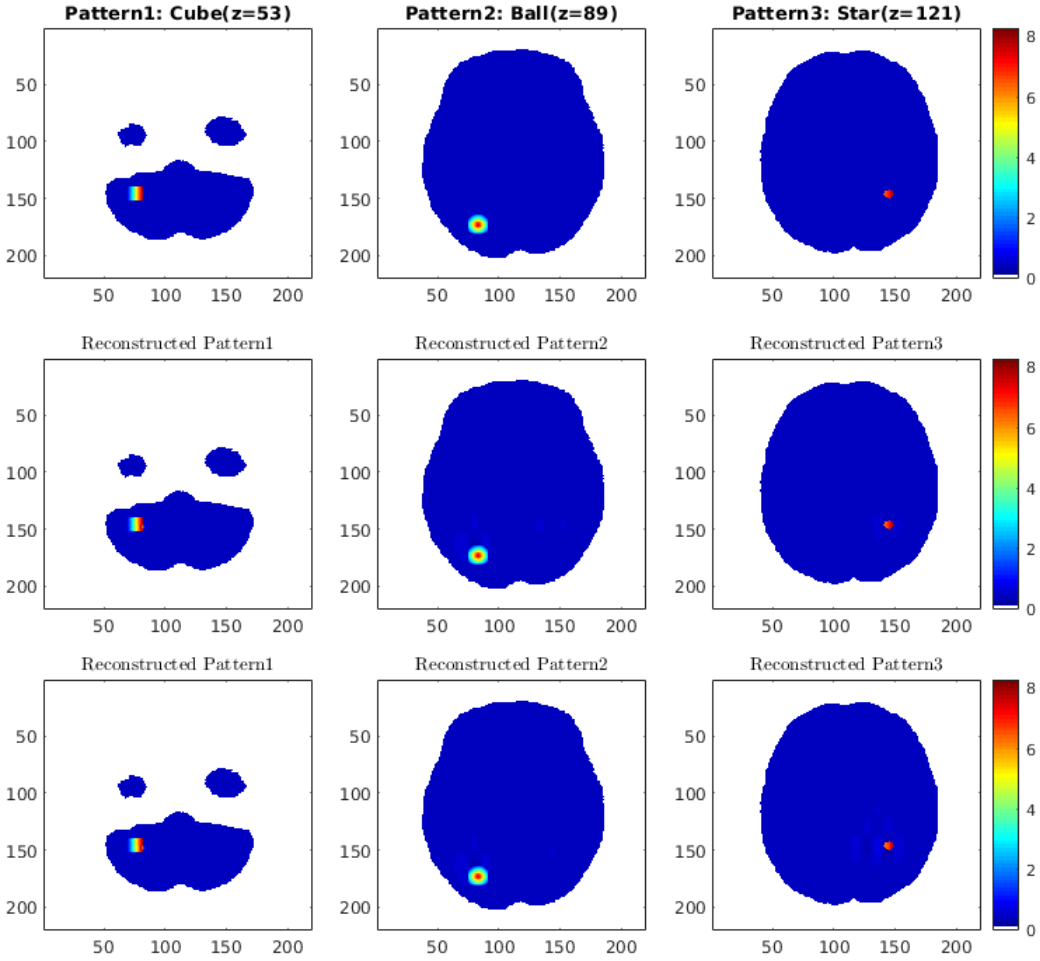

Figure 3.3: True and identified patterns of 3-D simulated data. First row: the true patterns of cube, ball and star shape at slice 53, 89 and 121; Second row: reconstructed patterns by mapping back testing result of wavelet-based Westfall-Young randomization test; Third row: reconstructed patterns by mapping back testing result of wavelet-based Holm's method. Color: value of the pattern.

nonparametric rank-sum test to obtain p-values. The testing result of the reconstructed contrast effect from basis-space testing is given in Figure3.4(a). We flag the region of difference by $\delta = 0.02$, a threshold suggested by Zhu et al. [83]. The region detected by point-wise test is given in Figure3.4(b). Point-wise test flagged some regions in very high emission wavelength. These regions are not flagged by using basis space test.

Our result reveals a major local region around excitations 330 - 420 nms and emissions 420 - 520 nms on the EEM that reflects differences between the normal and pre-cancer samples. In particular, normal samples tend to have higher intensity than the pre-cancer samples in this region. If desired, we may adjust the threshold to identify locations with other levels of differences. The flagged regions may serve as bio-markers for future disease assessment and diagnosis.

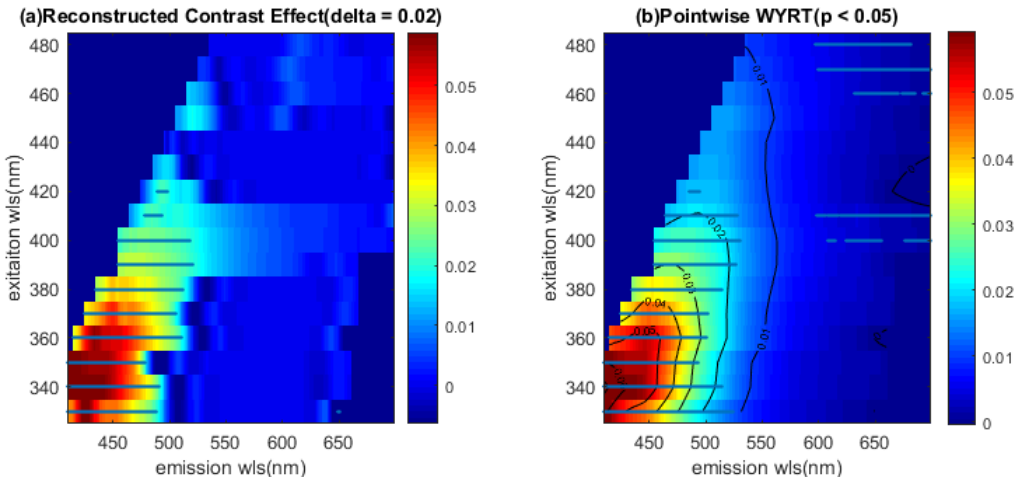

Figure 3.4: The identified region of difference between normal and pre-cancer groups based on EEM data. (a) Reconstructed mean difference after wavelet-based West-fall Young test: flagged region with reconstructed value larger than 0.02. (b) Region of difference identified by point-wise West-fall Young test with $\alpha = 0.05$

3.6.2 Tensor-based morphometry images of human brain

In the second example, we apply our method on the 3D Tensorbased morphometry (TBM) preprocessed MRI scans from the ADNI study. Details of the dataset is described in Section 1.2.4. We have images of 816 participants, among which, 118 are Alzheimer's Disease (AD) patients, 384 are with mild cognition impairment (MCI) and the rest 228 are normal (NL). One goal is to detect regions that reflect volumetric brain difference between AD vs. NL and between MCI vs. NL.

Wavelet transformation resulted in around 10^8 components. At each component, we perform a two-sample t-test. Based on the p-values, compression is performed and only tests with p-value smaller than the threshold $\tau = 1e-4$ are retained. We applied Westfall-Young randomization adjustment to the remaining p-values, and rejected hypotheses with adjusted p-values smaller than 0.05.

To identify the regions of difference, we first obtain the wavelet representation of the sample mean difference between two groups of interest. We keep the coefficients corresponding to the rejected tests and set the rest to be zero. We then map back the resulting sequence of coefficients to reconstruct the contrast effect. To demonstrate regions of difference, we plot zero as white and show the differences by color, with red indicating inflation and blue indicating atrophy. Figure 3.5 shows one slice of the brain. It shows that the main differences between AD and NL groups are in the middle part of the brain. Specifically, ventricles are enlarged while the surrounded part gets smaller.

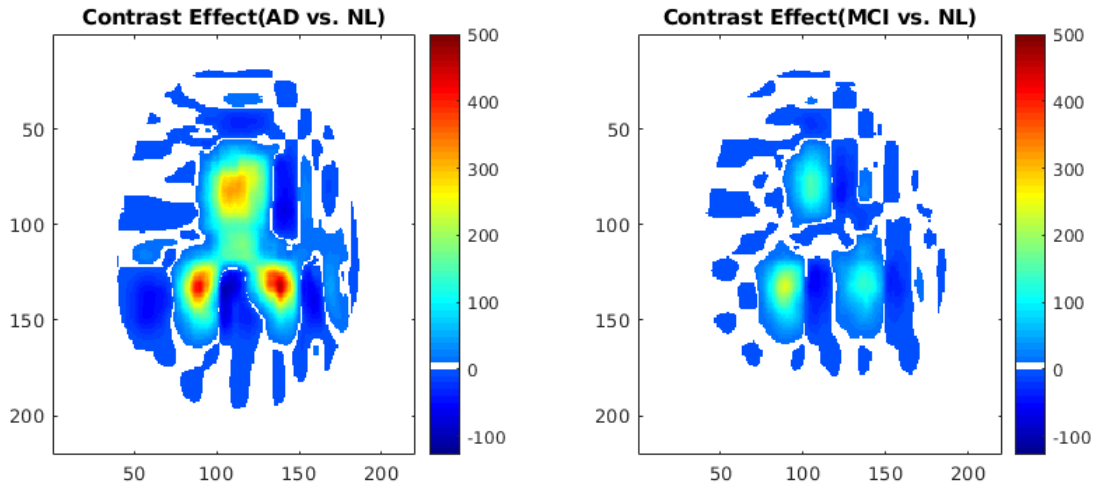

Figure 3.5: Analysis result of 3D tensor-based morphometry (TBM) data. TBM data measures the volumetric tissue differences relative to a standard template among 816 patients with Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), or normal controls. Color means the difference between the first group vs. the second group. Red indicates inflation (positive difference) and blue indicates atrophy (negative difference)

Chapter 4

Analysis of Functional Data Supported on Manifold

Motivated by monkeys' cortical measurements data described in Section 1.2.5, we consider the analysis of longitudinal functional data that are spatially dense, longitudinally sparse, and collected on irregular time grid. Unlike traditional functional data that are defined on domains in Euclidean space, the functional data we considered here are measured on brain cortex, which has more complex geometries such as sulcus and gyrus. To deal with functional data with such characteristics, we first map the cortical surface to a standard sphere as shown in Figure 1.8. With this mapping, we are able to use spherical wavelet to represent the data. This transforms that data to wavelet space. For each wavelet component, we use PACE (reviewed in Section 1.3.4) to estimate the trajectory from 0 to 36 months for for each monkey by borrowing information from all monkeys. With the help of PACE, we obtain estimated wavelet coefficients on any arbitrary time marker for every subject. The estimated trajectories for all wavelet components are then used in basis-space testing to identify significant spatial/temporal regions. Using the basis-space testing procedure proposed in Chapter 3, we are able to answer two questions related to the early development of monkeys' brains: (1) Are there any asymmetric patterns in cortical measurements between left and right brain? If so, when and where the asymmetries occur; (2) Are there significant increase/decrease in cortical measurements overtime? If so, where these trends occur. To be

specific, we only consider the cortical thickness measurement in our study.

The outline of the following sections is as follows: Section 4.1 describes our two analysis problems and provides details of the analysis procedures. Section 4.2 verifies the proposed analytic framework through a simulated study. Section 4.3 demonstrates the analysis results.

4.1 Problem setup and methods

Let $\mathbf{Y}^L(x,t) = \{y_i^L(x,t),\ i = 1,\ldots,N\}(x \in \mathcal{X}, t \in \mathcal{T})$ be the longitudinal thickness measurements on the left semi-cortical surface of all the monkeys, where i is the index of monkey, x is the coordinates on the semi-cortical surface and t is the time marker. Similarly, we define the thickness measurements of the right half brains by $\mathbf{Y}^R(x,t) = \{y_i^R(x,t),\ i = 1,\ldots,N\}$. $(x \in \mathcal{X}, t \in \mathcal{T})$, where $\mathcal{X}$ stands for the domain of semi-cortical surface and $\mathcal{T}$ stands for the domain of time. Based on these definitions, we denote the longitudinal thickness measurements on the whole cortical surface by $\mathbf{Y}(x,t) = \{y_i(x,t)\} = \{(y_i^L(x,t), y_i^R(x,t))\}$. Further, we denote the mean functions of $\mathbf{Y}^L(x,t)$, $\mathbf{Y}^R(x,t)$ and $\mathbf{Y}(x,t)$ by $\mu^L(x,t)$, $\mu^R(x,t)$ and $\mu(x,t)$ respectively. We establish our hypothesis in data domain as a family of hypotheses indexed by (x,t). Specifically, for the testing of asymmetric patterns, for fixed x and t, our hypothesis is:

$$
\begin{aligned}
H_0 :&\quad \mu^L(x,t) = \mu^R(x,t) \\
H_a :&\quad \mu^L(x,t) \neq \mu^R(x,t).
\end{aligned}
$$

For the testing of trends, for any fixed x and time markers t and t', our hypothesis is:

$$\begin{aligned} H_0: &\quad \mu(x,t) = \mu(x,t') \\ H_a: &\quad \mu(x,t) \neq \mu(x,t'). \end{aligned}$$

Suppose we have a standard semi-spherical surface $\mathcal{D}$, and there is an one-to-one map that projects y_i in $\mathcal{X}$ to $\tilde{y}_i$ in $\mathcal{D}$. On $\mathcal{D}$, we define a set of spherical wavelet bases, $\phi_k(d), k = 1, 2, \ldots, d \in \mathcal{D}$, such that

$$\tilde{y}_i(d,t) = \sum_k^{\infty} c_{i,k}(t)\phi_k(d),$$

where $c_{i,k}(t)$ is the wavelet coefficient corresponding to the kth wavelet component for the ith monkey at time t. Similarly, we have the wavelet representations for the left brain and right brain as follows.

$$\begin{aligned} \tilde{y}_i^L(d,t) &= \sum_k^{\infty} c_{i,k}^L(t)\phi_k(d), \\ \tilde{y}_i^R(d,t) &= \sum_k^{\infty} c_{i,k}^R(t)\phi_k(d). \end{aligned}$$

Let $u_k^L(t), u_k^R(t)$ and $u_k(t)$ be the mean of $\{c_{i,k}^L(t), i = 1, \ldots, N\}$, $\{c_{i,k}^R(t), i = 1, \ldots, N\}$, and $\{c_{i,k}(t), i = 1, \ldots, N\}$ respectively. Equivalently, on discrete time markers: $t_1, \ldots, t_S$, denote $\mathbf{u}^L(t_s) = (u_1^L(t_s), u_2^L(t_s), \ldots)$, $\mathbf{u}^R(t_s) = (u_1^R(t_s), u_2^R(t_s), \ldots)$, and $\mathbf{u}(t_s) = (u_1(t_s), u_2(t_s), \ldots)$. We establish the testing problem in wavelet domain as a family of hypotheses indexed by k and t. Specifically,

(1) at fixed time t, for testing of asymmetries, we set the hypotheses at time t by:

$$\begin{aligned} H_0 : &\quad \mathbf{u}^L(t) = \mathbf{u}^R(t) \\ H_a : &\quad \mathbf{u}^L(t) \neq \mathbf{u}^R(t),\ for\ some\ k \end{aligned}$$

(2) at two different time markers t and t', for testing of trends, the hypotheses are claimed as:

$$\begin{aligned} H_0 : &\quad \mathbf{u}(t) = \mathbf{u}(t') \\ H_a : &\quad \mathbf{u}(t) \neq \mathbf{u}(t'),\ for\ some\ k \end{aligned}$$

Our data consist of realizations of $\mathbf{Y}(x,t)$ for 36 monkeys at different time grids. Denote the time grids for the ith monkey by $t_{i1}, \ldots, t_{iN_i}$, where N_i is the number of measurements for the ith monkey. After mapping onto a standard semi-sphere, we have realizations of $\tilde{\mathbf{Y}}(d,t)$ on $t_{i1}, \ldots, t_{iN_i}$. Performing spherical wavelet decomposition on $\tilde{\mathbf{Y}}(d,t)$, we express data as linear combinations of a set of basis functions $\phi_k(d)(k = \{1, 2, \ldots\})$:

$$\begin{aligned} \tilde{Y}_i^L(d,t_j) &= \sum_k c_{i,k}^L(t_j)\phi_k(d) \\ \tilde{Y}_i^R(d,t_j) &= \sum_k c_{i,k}^R(t_j)\phi_k(d) \\ \tilde{Y}_i(d,t_j) &= \sum_k c_{i,k}(t_j)\phi_k(d), \end{aligned}$$

where $c_{i,k}(t_j)$ is the kth coefficient for the ith monkey at time t ($i = 1, \ldots, 36$, $j = 1, \ldots, N_i$). To perform the two tests established above, we need to have evaluations of the wavelet coefficients on the same set of time markers. For that purpose, we estimate the whole

trajectory of every wavelet component for each monkey using PACE (reviewed in Section 1.3.4). In Figure 4.1, we show the trajectories of the first five wavelet components for the monkey #1.

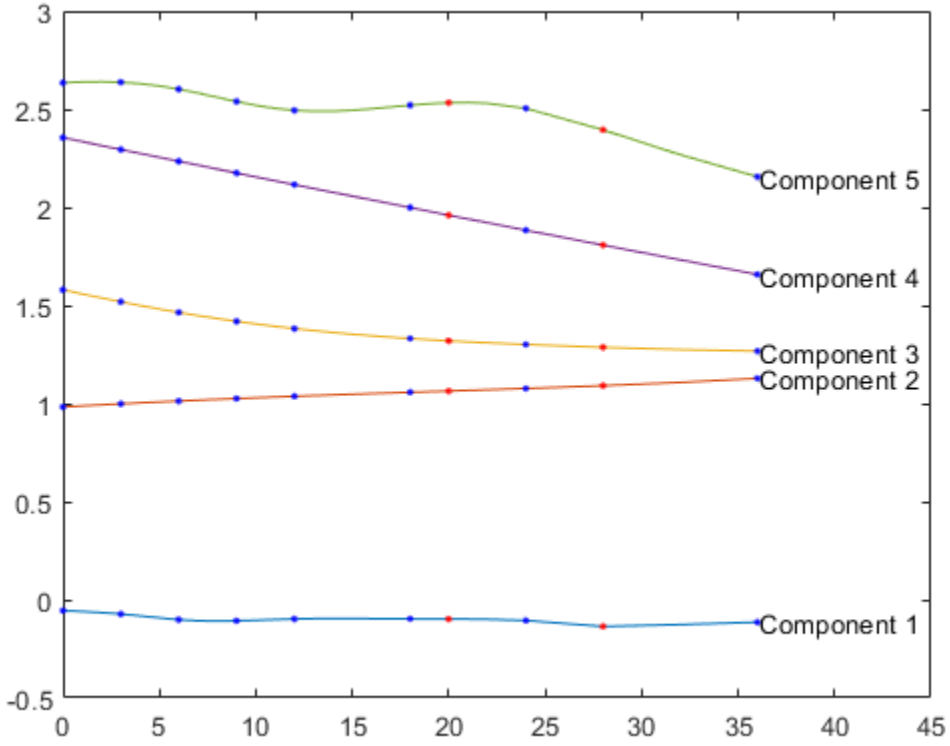

Figure 4.1: Component-wise trajectory (of Monkey #1) obtained by using PACE. The trajectories of the first five spherical wavelet components from 0 to 36 months. Red dots are the real observations, and blue dots indicate the selected time markers, on which we detect trends

Because the brains are typically more dynamic in the first year than in the third year, to identify the asymmetric patterns between left and right brain across time, we select 0, 3, 6, 9, 12, 18, 24 and 36 months as the time markers on which to test asymmetries. In addition, we will perform paired t-test to test difference between left and right brains at each wavelet component, as we are comparing the right and left brain within each individual. To identify

the increase/decrease trends, we select the time markers to be 0, 3, 6, 9, 12, 15, 18, 21, 24, 28, 32, and 36 months. We apply paired t-test on the estimated trajectories. After testing, we will keep the rejected components while set the rest components to zero. The resulting sequence of wavelet coefficients will be transformed back into data domain and find the regions of interest.

4.2 Simulation study

In this section, we will evaluate the analysis procedure proposed in the previous section through a simulated example. To mimic the real data, we first generate a reference image on the standard semi-sphere at 0 month and then add increasing/decreasing trends onto it. From the mean left-vs-right contrast at month 0, we obtain the reference image by setting values with absolute values smaller than 0.3 to zero, The generated true asymmetric patterns at selected time markers are shown in Figure 4.2

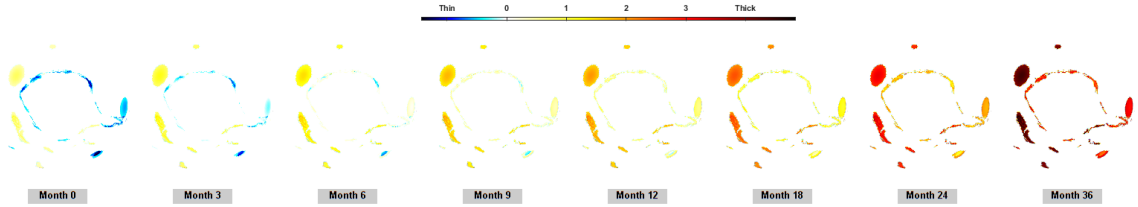

Figure 4.2: True simulated asymmetric patterns at selected time markers: 0, 3, 6, 9, 12, 18, 24, 36 month. Color indicates value of left-vs-right contrast: red represents left-thicker-than-right regions and blue represents right-thicker-than-left-regions. The smallest value of difference is 0.3.

We then add random noise based on a N(0,0.01) distribution. In the longitudinal direction, we use the same sampling time markers as the real data to mimic the sparseness pattern. On the simulated data, the analysis procedure is performed. We first conduct spherical wavelet decomposition and use PACE to obtain the estimated trajectories on each

wavelet component. Based on the estimated trajectories, we perform basis-space testing with Westfall-Young randomization adjustment. The threshold used for compression is set to be 1e-4. In data domain, regions with reconstructed contrast effects greater than 0.3 are flagged to be significant, and based on which we calculated FDR, Empirical Power and specificity. The means of these three statistics are 0.0671, 0.9735 and 0.9983 respectively based on 50 simulations. When transforming the testing results back into data domain, we achieve the reconstructed left-vs-right contrast at the selected time markers shown in Figure 4.3.

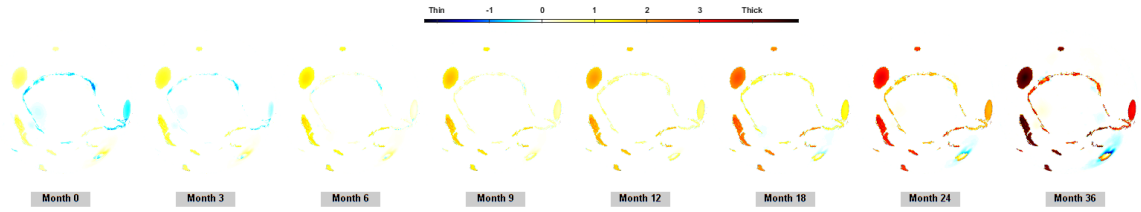

Figure 4.3: Reconstructed asymmetric patterns from testing results. Color indicates value of left-vs-right contrast: red represents left-thicker-than-right regions and blue represents right-thicker-than-left regions.

Similarly, we simulated data with longitudinal trend. Rather than generating data on both halves of brains, we generated data on the standard semi-sphere using the same reference image. In this simulation, we select more time-markers for the testing of trends. The simulated increasing/decreasing patterns are shown in Figure 4.4.

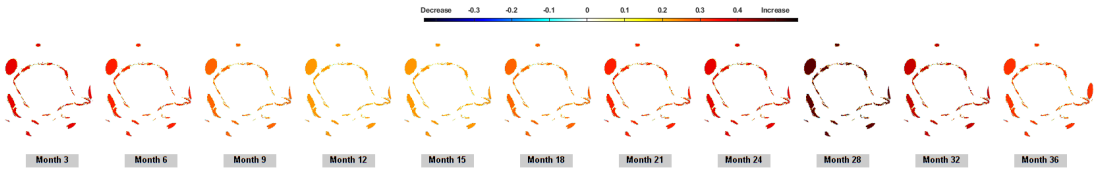

Figure 4.4: True trends of the simulated data (difference between two coherent time markers). Color indicates the difference between the latter time marker vs the former time marker. Red represents increasing regions and blue represents decreasing regions.

On the simulated pattern, we add i.i.d random noise based on a $N(0, 0.01)$ distribution. In the longitudinal direction, we use the same sampling time markers as the real data to

mimic the sparseness pattern. The analysis procedure stays the same as that for the testing of asymmetries. In the data domain, the means of the summary statistics, FDR, empirical power and specificity are 0.0474, 0.9720 and 0.9980 respectively based on 50 simulations. Based on results from one simulation, we plot the reconstructed trends (differences between month 3 vs. month 0, month 6 vs. month 3, and so on) in data domain in Figure 4.5

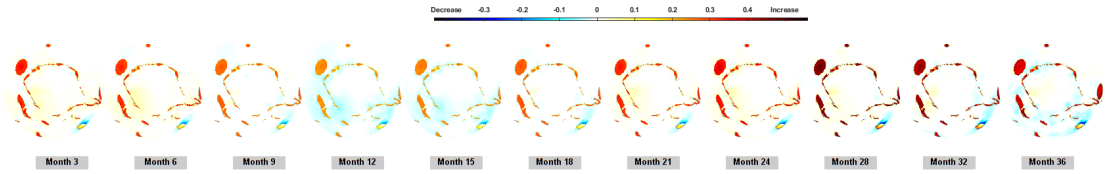

Figure 4.5: Reconstructed trends from testing results based on simulated data: Red represents becoming-thicker regions and blue represent becoming-thinner regions.

Our simulated examples mimic the real data, in terms of the manifold geometry, dimension, longitudinal sparseness patterns. The simulations demonstrate that our analysis procedure is able to successfully identify the regions of interests.

4.3 Results

In this section, we will demonstrate results for the cortical thickness data analysis. Following the descriptions of major areas in macaque monkey's brain in Seidlitz et al. [59], we are able identify clustered regions that are significantly different between left and right brain, and along the temporal direction.

Asymmetric patterns in cortical thickness are demonstrated on reconstructed left-vs-right contrast surface. These contrast surfaces are obtained by transforming significant components in the mean differences back to the data domain. Results are demonstrated in Figure 4.6. From Figure 4.6, we observe that on both the dorsal and medial surfaces, in most re-

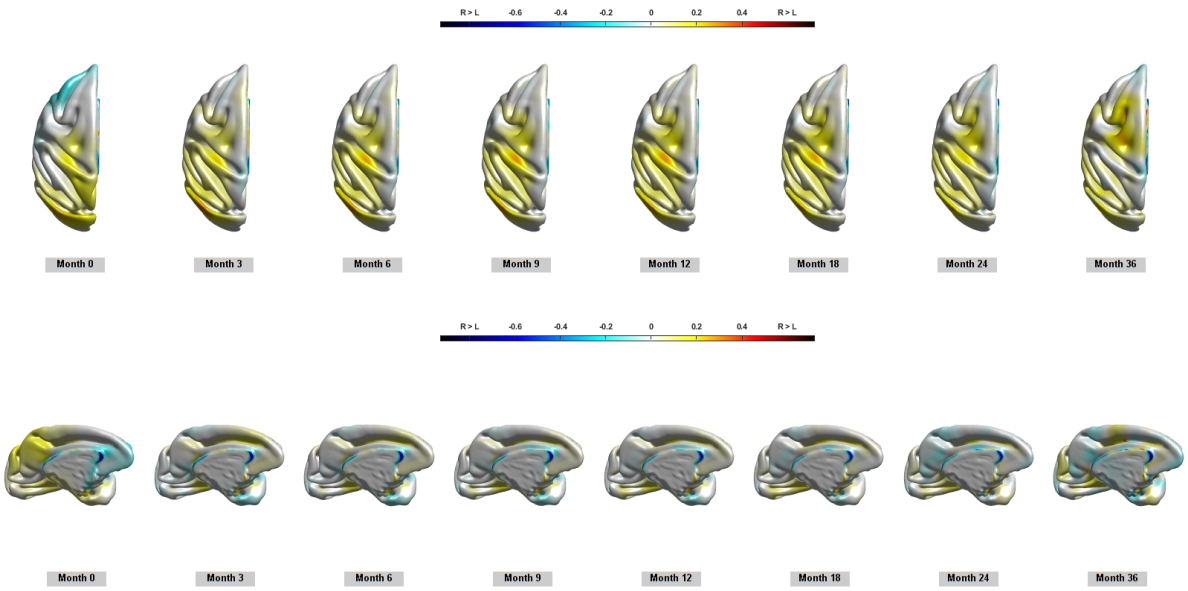

Figure 4.6: Analysis results of asymmetric patterns in cortical thickness. On both dorsal surface (top line) and medial surfaces (bottom line), red represents left-thicker-than-right regions and blue represents right-thicker-than-left regions.

gions, the left brain has thicker cortical thickness than the right, especially in the actuate sulcus, central sulcus and lunate sulcus on the dorsal surface, as well as the temporal and cingulate cortex on the medial surface.

On the dorsal surface, the leftward asymmetry in a small portion in central sulcus appears from the 3rd month, expands till the 18th month and gradually shrinks afterwards and finally disappears at 36 months. The leftward asymmetry in arcuate sulcus is relatively small in the first 3 months, and gradually expends till the 36th month. A small portion in lunate sulcus in left brain is thicker than right brain at the very beginning. The asymmetry expands till 18 months and shrinks afterwards.

On the medial surface, we see three major significant clusters consistently identified from 0 to 36 months, including the leftward asymmetries in the medial temporal cortex and posterior cingulate sulcus and the rightward asymmetry in the cingulate gyrus. The leftward

asymmetry also appears in a small portion in the medial orbitofrontal cortex.

Changing patterns in macaque monkey's brain over time are plotted on the reconstructed mean contrast surface between every two consecutive time markers in Figure 4.7.

On the lateral surface of left brain, we see a constant increase in the lateral sulcus in the first two years. Afterwards, this trend reduces and a small portion in this region shrinks in the last four months of the third year. Except this region, the rest parts get thinner in the first year. From the second year, the cortical thickness increases in the central sulcus. The increasing region gradually expands and includes arcuate sulcus from the end of the second year. The cortical thickness keeps decreasing in major regions in inferior occipital sulcus and lunate sulcus till 18 months. The decreasing regions shrink and finally disappear from the 32nd month.

The frontal and ventral views of the whole brain are shown in Figure 4.7. From Figure 4.7, we observe that in general, the right brain shows more dynamic changes than the left brain. Evidently, in the central sulcus, the right brain gets thicker in this area gradually from 0 month till 24 months. In the third year, a dramatic growth is found in the central sulcus in right brain. The same area of left brain shrinks in the first 15 months. Additionally, a small portion in this area gets thicker. It expands gradually and finally includes a major proportion of the frontal surface. On the ventral surface, we see a significant increase in the right half in the first 4 months. The increase slows down and finally disappear at the 24th month.

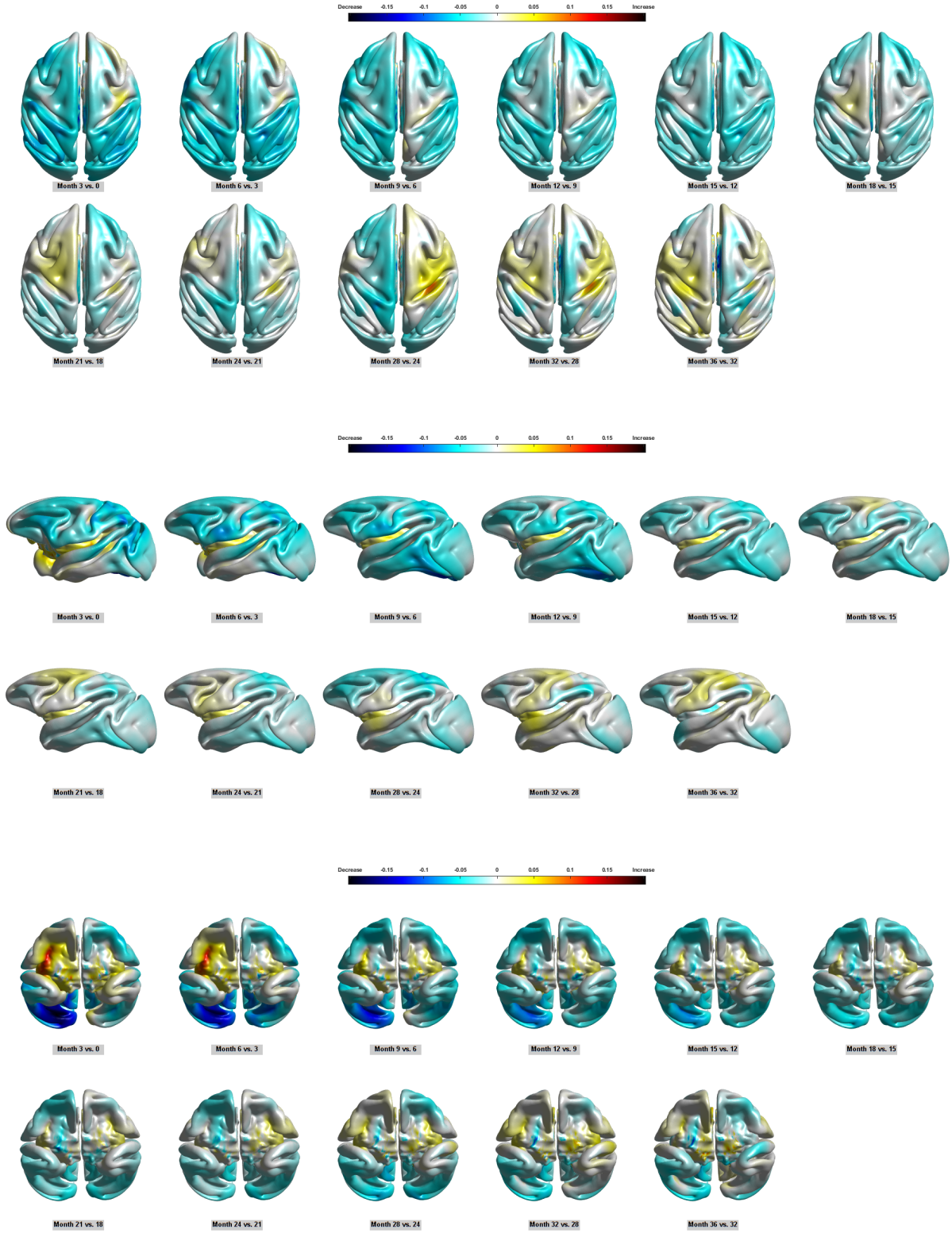

Figure 4.7: Analysis results of changing patterns in cortical thickness. From the frontal view (top line), the lateral view (middle line), and the ventral view (bottom line), red represents increases and blue represents decreases.

Chapter 5

Conclusion and Discussion

This dissertation focuses on developing novel estimation, inference and analysis methods for data arising from complex systems.

In Chapter 2, we proposed two mutation rate estimators for the birth-death process model and the generalized birth-death process model. To our knowledge, this is the first approach that is available to estimate a time-varying mutation rate in fluctuation analysis. As a likelihood-free method, ABC provides us with a flexible tool, that goes beyond the traditional likelihood-based approaches. It allows us to consider more complex simulation models with more structures such as cell death and multiple fission. Through simulation, we showed that, the new estimator provides comparable or even better accuracy than the traditional maximum likelihood estimator. We also illustrated that the ABC-GBD provides posterior samples of the joint distribution. With these posterior samples, it is straightforward to discover multi-modal behaviors, or characterize uncertainty of the model parameters. In the real data example, we showed that our method works well for the cases when the simulation model is computationally expensive. This is due to the fact that the GP surrogate model can dramatically reduce the number of callings of the simulator. Finally, our approach is equipped with an uncertainty control step which adaptively adjusts the number of training samples needed for the GPS and leads to improved mixing.

Beyond the two cases we considered in this dissertation, the newly proposed ABC framework can be adapted to accommodate more complicated forms of mutation rate function than the

constant or piece-wise constant function. For example, the mutation rate function could be a linear function or in some other parametric forms. The proposed ABC-GBD can be used to estimate all the underlying parameters simultaneously. It is also suitable for problem with higher dimensionality. Even when the dimension of parameter space is moderately high, the GP surrogate model may still perform well with moderate number of training samples with the help of computer experiment design techniques.

Based on the GPS-ABC framework, for complex system with functional outputs, we proposed a general simulation-based approach called wABC to estimate the underlying parameters. The proposed method relies on simulation models to estimate the parameters of interest, which avoids the difficulty of specifying the likelihood. We accommodate high-dimensional functional data by combining wavelet decomposition with compression, and achieve scalable computation using a Gaussian process surrogate to the simulator.

The proposed wABC approach is generally applicable to a large family of estimating problems incurred by complex systems, such as solving differential equations, and estimating parameters of a biological system. However, it requires a "simulator" to generate psudo-data. The simulator needs to describe the real system with sufficient accuracy. Otherwise, even if the wABC is tuned to perform well with simulated data, it may fail on real-word data.

While we have focused on systems with highly-stochastic functional outputs. The proposed framework is also suitable for deterministic systems in which the simulator yields a deterministic functional output given an input parameter. For example, many complex systems can be described by differential equations in which no random variables appear. In these situations, we just need to set $n = 1$ and $m = 1$ in wABC. These problems are often easier to solve than the stochastic systems considered here.

In Chapter 3, we proposed a basis-space testing procedure for region detection on functional data. Specifically, we showed that the procedure with randomization-based adjustment, Westfall-Young method, can strongly control FWER at a given significant level. Furthermore, under appropriate conditions, the Westfall-Young adjustment is asymptotically optimal when the number of tests goes to infinity. Furthermore, we showed that appropriate compression in basis domain can further improve the empirical power. These theoretical properties are verified by two simulation studies — one 1-d case and one 3-d case. Through simulation studies, we also showed that the proposed method is effective in identifying significant regions in data domain.

As a general approach, the advantages of our approach lies in several aspects: (i) We can adopt different types of bases. For the purpose of identifying significant regions, compactly supported basis is needed. (ii) There's no constraint on individual test used in basis space. This allows us to choose appropriate testing approach for un-adjusted p-values based on the data. (iii) We have proposed a general testing framework, which is not limited to one-sample or two-sample tests. (iv) It incorporates different multiple testing adjustment methods, including those that control FWER and those that control FDR. (v) We are able to detect regions of interest by transforming testing result in basis domain back to data domain.

In Chapter 4, we studied cortical hemispheric asymmetric and developing patterns in macaque monkeys during the early development stages. We proposed an analytic procedure for detecting significant cortical regions in macaque monkey's brain across time. The data are spatially dense, longitudinally sparse and collected on irregular time grids. By applying the proposed testing procedure to the estimated wavelet components at multiple selected time markers, we answered the following two questions: (1) Are there any increases/decreases in cortical thickness during the monkeys' early development? If so, where do these differences

occur? (2) Are there any asymmetric patterns in cortical thickness between left and right hemispheres? If so, when and where do the asymmetries occur?

For the first question, we find major leftward asymmetries in the artuate, central and lunate sulcus on the dorsal surface. In general, the asymmetric region is relatively small in the first three months, then expands till the second half of the second year and remains the same or shrinks afterwards. Existing studies are rare on cortial thickness asymmetric patterns. Scott et al. [58] has recently found leftward asymmetries of cortical volume in the frontal and parietal cortices in macaque monkeys from 1 week to 52 months, which is consistent with the leftward asymmetries we discovered in our study. On the medial surface, we saw mainly three significant clusters, including the leftwards asymmetries in the medial temporal cortex and posterior cingulate sulcus and the rightward asymmetry in the cingulate gyrus. The discoveries of the leftward asymmetries are consistent with the findings in human brains from birth to 2 years old [31].

For the second question, we saw that more dynamic changes in the right brain than in the left brain. For both halves, in the first year, the thickness in the central part increases while the rest decreases. In the following years, the increasing region expands and covers the frontal cortex.

Studying macaque brains is important to understand human brains and provides us with reference in the studies of neuro-developmental disorders. However, the high dimensionality and sparseness of the data make the traditional method intractable. The data is too sparse and most imputation approach fail. Furthermore, point-wise tests will lose power for such large number of tests. Our basis-space testing approach shows great advantages over the traditional methods in several aspects: (1) It is shown to be more powerful and scalable than the traditional point-wise test. (2) As the regions are detected by reconstructing contrast effect through inversely transforming significant components into the data domain,

the reconstructed image provides not only information of statistical significance but also the magnitude of differences, which is more intuitively interpretable. In the future, this method could be applied to other cortical measurements, such as cortical surface area, and sulcal depth to provide more comprehensive understanding of macaque brains in the early development stages.

Bibliography

[1] (2014), "GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation," *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*.

[2] Asteris, G. and Sarkar, S. (1996), "Bayesian procedures for the estimation of mutation rates from fluctuation experiments," *Genetics*, 142, 313–326.

[3] Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

[4] Binois, M., Gramacy, R. B., and Ludkovski, M. (2018), "Practical heteroskedastic Gaussian process modeling for large simulation experiments," *Journal of Computational and Graphical Statistics*, 1–41.

[5] Bowman, J., Senior, T., and Uslenghi, P. (1987), *Electromagnetic and acoustic scattering by simple shapes*, New-York: Hemisphere Publishing Corporation.

[6] Bridwell, D. A., Cavanagh, J. F., Collins, A. G. E., Nunez, M. D., Srinivasan, R., Stober, S., and Calhoun, V. D. (2018), "Moving Beyond ERP Components: A Selective Review of Approaches to Integrate EEG and Behavior," *Frontiers in Human Neuroscience*, 12, 106.

[7] Brown, P. J., Vannucci, M., and Fearn, T. (1998), "Bayes Model averaging with selection of regressors," *Journal of the Royal Statistical Society, Series B*, 60, 627–641.

[8] Cabaña, A., Estrada, A. M., Peña, J., and Qurioz, A. J. (2017), "Permutation tests in the two-sample problem for functional data," in *Functional Statistics and Related Fields*,

eds. Aneiros, G., G. Bongiorno, E., Cao, R., and Vieu, P., Cham: Springer International Publishing, pp. 77–84.

[9] Cao, G., Yang, L., and Todem, D. (2012), "Simultaneous inference for the mean function based on dense functional data," *Journal of Nonparametric Statistics*, 24, 359–377.

[10] Cardot, H. (2005), "Nonparametric regression for functional responses with application to conditional functional principle component analysis," *Available online: http://www.lsp.ups-tlse.fr/Recherche/Publications/2005/car01.pdf*.

[11] Chang, C., Lin, X., and Ogden, R. T. (2017), "Simultaneous confidence bands for functional regression models," *Journal of Statistical Planning and Inference*, 188, 67 – 81.

[12] Chiou, J., Müller, H., and Wang, J. (2003), "Functional quasi-likelihood regression models with smooth random effects," *Journal of the Royal Statistical Society, Series B*, 65, 405–423.

[13] Cox, D. D. and Lee, J. S. (2008), "Pointwise testing with functional data using the Westfall–Young randomization method," *Biometrika*, 95, 621–634.

[14] Degras, D. A. (2011), "Simultaneous Confidence Bands for Nonparametric Regression with Functional Data," *Statistica Sinica*, 21, 1735–1765.

[15] Demerec, M. (1945), "Production of Staphylococcus strains resistant to various concentrations of penicillin," *Proceedings of the National Academy of Sciences*, 31, 16–24.

[16] Didelot, X., Everitt, R. G., Johansen, A. M., Lawson, D. J., et al. (2011), "Likelihood-free estimation of model evidence," *Bayesian analysis*, 6, 49–76.

[17] Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis*, Springer-Verlag.

[18] Geweke, J. (1992), "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," *Bayesian Statistics 4*.

[19] Ghale-Joogh, H. S. and Hosseini-Nasab, S. M. E. (2018), "A two-sample test for mean functions with increasing number of projections," *Statistics*, 52, 852–873.

[20] Guo, J., Zhou, B., and Zhang, J.-T. (2018), "Testing the equality of several covariance functions for functional data: A supremum-norm based test," *Computational Statistics & Data Analysis*, 124, 15 – 26.

[21] Hall, P. and Keilegom, I. V. (2007), "TWO-SAMPLE TESTS IN FUNCTIONAL DATA ANALYSIS STARTING FROM DISCRETE DATA," *Statistica Sinica*, 17, 1511–1531.

[22] Horváth, L. and Kokoszka, P. (2012), *Inference for Functional Data with Applications*, Springer Verlag.

[23] Hua, X., Leow, A. D., Parikshak, N., Lee, S., Chiang, M.-C., Toga, A. W., Jack Jr, C. R., Weiner, M. W., Thompson, P. M., Initiative, A. D. N., et al. (2008), "Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects," *Neuroimage*, 43, 458–469.

[24] James, G. M., Hastie, T. J., and Sugar, C. A. (2000), "Principal component models for sparse functional data," *Biometrika*, 87, 587–602.

[25] James, G. M. and Sugar, C. A. (2003), "Clustering for sparsely sampled functional data," *Journal of the American Statistical Association*, 98, 397–408.

[26] James, G. M., Wang, J., and Zhu, J. (2009), "Functional linear regression that's interpretable," *The Annals of Statistics*, 2083–2108.

[27] Jiang, Q., Meintanis, S. G., and Zhu, L. (2017), "Two-sample tests for multivariate functional data," in *Functional Statistics and Related Fields*, eds. Aneiros, G., G. Bongiorno, E., Cao, R., and Vieu, P., Cham: Springer International Publishing, pp. 145–154.

[28] Jones, M., Wheldrake, J., and Rogers, A. (1993), "Luria-Delbrück fluctuation analysis: estimating the Poisson parameter in a compound Poisson distribution," *Computers in biology and medicine*, 23, 525–534.

[29] Koltchinskii, V. and Minsker, S. (2013), "L_1-Penalization in Functional Linear Regression with Subgaussian Design," *arXiv preprint arXiv:1307.8137*.

[30] Lea, D. E. and Coulson, C. A. (1949), "The distribution of the numbers of mutants in bacterial populations," *Journal of genetics*, 49, 264.

[31] Li, G., Wang, L., Shi, F., Gilmore, J. H., Lin, W., and Shen, D. (2015), "Construction of 4D high-definition cortical surface atlases of infants: Methods and applications," *Medical image analysis*, 25, 22–36.

[32] Liao, X., Yan, X., Xia, W., and Luo, B. (2010), "A fast optimal Latin hypercube design for Gaussian process regression modeling," in *Advanced Computational Intelligence (IWACI), 2010 Third International Workshop on*, IEEE, pp. 474–479.

[33] Lin, Z., Cao, J., Wang, L., and Wang, H. (2015), "A Smooth and Locally Sparse Estimator for Functional Linear Regression via Functional SCAD Penalty," *arXiv preprint arXiv:1510.08547*.

[34] Liu, J. and Ye, J. (2010), "Moreau-Yosida Regularization for Grouped Tree Structure Learning," in *Advances in Neural Information Processing Systems 23*, eds. Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., Curran Associates, Inc., pp. 1459–1467.

[35] Lu, B., Castillo, I., Chiang, L., and Edgar, T. F. (2014), "Industrial PLS model variable selection using moving window variable importance in projection," *Chemometrics and Intelligent Laboratory Systems*, 135, 90 – 109.

[36] Luria, S. E. and Delbrück, M. (1943), "Mutations of bacteria from virus sensitivity to virus resistance," *Genetics*, 28, 491.

[37] Ma, S., Yang, L., and Carroll, R. J. (2012), "A Simultaneous confidence band for sparse longitudinal regression," *Statistica Sinica*, 22, 95 – 122.

[38] Marin, J.-M., Pudlo, P., Robert, C. P. R., and Ryder, R. J. (2012), "Approximate Bayesian computational methods," *Stat. Comput.*, 22, 1167–1180.

[39] Marín, N., MacKinnon, N. B., MacAulay, C. E., Chang, S. K., Atkinson, E. N., Cox, D. D., Serachitopol, D., Pikkula, B. M., Follen, M., and Richards-Kortum, R. R. (2006), "Calibration standards for multicenter clinical trials of fluorescence spectroscopy for in vivo diagnosis," *Journal of biomedical optics*, 11, 014010.

[40] McKay, M. D., Beckman, R. J., and Conover, W. J. (1979), "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, 21, 239–245.

[41] Meeds, E. and Welling, M. (2014), "GPS-ABC: Gaussian process surrogate approximate Bayesian computation," *arXiv preprint arXiv:1401.2838*.

[42] Meinshausen, N., Maathuis, M. H., Bühlmann, P., et al. (2011), "Asymptotic optimality of the Westfall–Young permutation procedure for multiple testing under dependence," *The Annals of Statistics*, 39, 3369–3391.

[43] Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P., and Morris, J. S. (2015), "Bayesian function-on-function regression for multilevel functional data," *Biometrics*.

[44] Miller, F., Vandome, A., and McBrewster, J. (2010), *Curse of Dimensionality*, VDM Publishing House Ltd.

[45] Morris, J. S. (2015), "Functional Regression," *Annual Review of Statistics and Its Application*, 2, 321–359.

[46] Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., and Coombes, K. R. (2008), "Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models," *Biometrics*, 64, 479–489.

[47] Ono, M., Kubik, S., and Abernathey, D. (1990), "Atlas of the cerebral sulci Georg," .

[48] Pini, A. and Vantini, S. (2016), "The Interval Testing Procedure: A General Framework for Inference in Functional Data Analysis," *Biometrics*, 72, 835–845.

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

[50] Pomann, G.-M., Staicu, A.-M., and Ghosh, S. K. (2016), "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.

[51] Pritchard, J. K., T., S. M., A., P.-L., and W., F. M. (1999), "Population growth of human Y chromosomes: a study of Y chromosome microsatellites," *Mol. Biol. Evol.*, 16, 1791–1798.

[52] Ramsay, J. O. and Li, X. (1998), "Curve registration," *J. Royal Statist. Soc. Ser. B*, 60, 351–363.

[53] Ramsay, J. O. and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer-Verlag.

[54] Rice, J. A. and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society, Series B*, 53, 233–243.

[55] Sadegh, M. and Vrugt, J. A. (2014), "Approximate bayesian computation using markov chain monte carlo simulation: Dream (abc)," *Water Resources Research*, 50, 6767–6787.

[56] Sapatinas, T. and Paparoditis, E. (2016), "Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data," *Biometrika*, 103, 727–733.

[57] Scheipl, F., Staicu, A.-M., and Greven, S. (2014), "Functional Additive Mixed Models," *Journal of Computational and Graphical Statistics*, 24, 477–501.

[58] Scott, J. A., Grayson, D., Fletcher, E., Lee, A., Bauman, M. D., Schumann, C. M., Buonocore, M. H., and Amaral, D. G. (2016), "Longitudinal analysis of the developing rhesus monkey brain using magnetic resonance imaging: birth to adulthood," *Brain Structure and Function*, 221, 2847–2871.

[59] Seidlitz, J., Sponheim, C., Glen, D., Frank, Q. Y., Saleem, K. S., Leopold, D. A., Ungerleider, L., and Messinger, A. (2018), "A population MRI brain template and analysis tools for the macaque," *Neuroimage*, 170, 121–131.

[60] Smaga, Ł. and Zhang, J.-T. (2018), "Linear Hypothesis Testing With Functional Data," *Technometrics*, 0, 1–12.

[61] Staniswalis, J. G. and Lee, J. J. (1998), "Nonparametric regression analysis of longitudinal data," *Journal of the American Statistical Association*, 93, 1403–1418.

[62] Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen,

A. N., Corneveaux, J. J., DeChairo, B. M., Potkin, S. G., Weiner, M. W., and Thompson, P. M. (2010), "Voxelwise genome-wide association study (vGWAS)," *NeuroImage*, 53, 1160 – 1174, imaging Genetics.

[63] Tang, R. and Müller, H.-G. (2008), "Pairwise curve synchronization for functional data," *Biometrika*, 95, 875.

[64] Turner, B. M. and Van Zandt, T. (2012), "A tutorial on approximate Bayesian computation," *Journal of Mathematical Psychology*, 56, 69–85.

[65] Vanderelst, D., Steckel, J., Boen, A., Peremans, H., and Holderied, M. W. (2016), "Place recognition using batlike sonar," *Elife*, 5, e14188.

[66] Vannucci, M. and Corradi, F. (1999), "Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 971–986.

[67] Vsevolozhskaya, O., Greenwood, M., and Holodov, D. (2014), "Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis," *Ann. Appl. Stat.*, 8, 905–925.

[68] Wang, J.-L., Chiou, J.-M., and Müller, H. (2015), "Review of functional data analysis," *Annu. Rev. Statist.*, 1–41.

[69] Wegmann, D., Leuenberger, C., and Excoffier, L. (2009), "Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood," *Genetics*.

[70] Westfall, P. H. and Young, S. S. (1993), "Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment (Wiley Series in Probability and Statistics)," .

[71] White, T., Su, S., Schmidt, M., Kao, C.-Y., and Sapiro, G. (2010), "The development of gyrification in childhood and adolescence," *Brain and cognition*, 72, 36–45.

[72] Wikipedia (2017), "Sonar — Wikipedia, The Free Encyclopedia," [Online; accessed 20-May-2017].

[73] Wu, X., Strome, E. D., Meng, Q., Hastings, P. J., Plon, S. E., and Kimmel, M. (2009), "A robust estimator of mutation rates," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 661, 101–109.

[74] Wu, X. and Zhu, H. (2015), "Fast maximum likelihood estimation of mutation rates using a birth–death process," *Journal of theoretical biology*, 366, 1–7.

[75] Yang, C., He, Z., and Yu, W. (2009), "Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis," *BMC Bioinformatics*, 10, 4.

[76] Yang, J., Zhu, H., Choi, T., and Cox, D. D. (2016), "Smoothing and Mean Covariance Estimation of Functional Data with a Bayesian Hierarchical Model," *Bayesian Anal.*, 11, 649–670.

[77] Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, 100, 577–590.

[78] Zheng, Q. (1999), "Progress of a half century in the study of the Luria–Delbrück distribution," *Mathematical biosciences*, 162, 1–32.

[79] Zheng, Q. (2002), "Statistical and algorithmic methods for fluctuation analysis with SALVADOR as an implementation," *Mathematical biosciences*, 176, 237–252.

[80] Zhou, J., Wang, N.-Y., and Wang, N. (2013), "Functional linear model with zero-value coefficient function at sub-regions," *Statistica Sinica*, 23, 25.

[81] Zhu, H., Brown, P. J., and Morris, J. S. (2011), "Robust, adaptive functional regression in functional mixed model framework," *J. Am. Statist. Ass.*, 495, 1167–1179.

[82] Zhu, H., Caspers, P., Morris, J. S., Wu, X., and Müller, R. (2018), "A Unified Analysis of Structured Sonar-Terrain Data Using Bayesian Functional Mixed Models," *Technometrics*, 60, 112–123.

[83] Zhu, H., Morris, J. S., Wei, F., and Cox, D. D. (2017), "Multivariate functional response regression, with application to fluorescence spectroscopy in a cervical pre-cancer study," *Computational statistics & data analysis*, 111, 88–101.

[84] Zhu, H., Versace, F., Cinciripini, P. M., Rausch, P., and Morris, J. S. (2018), "Robust and Gaussian spatial functional regression models for analysis of event-related potentials," *NeuroImage*, 181, 501 – 512.

Appendices

Appendix A

Appendix

A.1 Basis-space testing with Westfall-Young Randomization Adjustment

Step 1. Obtain the basis representation of observation $\{y_{ij}(x);\, i = 1, \ldots, g;\, j = 1, \ldots, n_i\}$ using a common basis set $\{\phi_k(x)\}_{k\in\{1,\ldots,K\}}$. Denote the coefficients$\{c_{ijk}; i = 1, \ldots, g;\, j = 1, \ldots, n_i; k = 1, \ldots, K\}$. This can be done by taking the inner-product (for orthonormal basis) or minimizing an objective function (for basis such as B-splines).

Step 2. Conduct a test $H_{0k} : \mathbf{A}\,\mathbf{u_k} = \mathbf{0}$ v.s. $H_{ak} : \mathbf{A}\,\mathbf{u_k} \neq \mathbf{0}$ using all coefficients $\{c_{ijk}\}$ for each fixed k. Denote the corresponding p-values $\{p_k;\, k = 1, ..., K\}$.

Step 3. Sort the p-values in increasing order: $p_{(1)} \leq p_{(2)} \leq ...p_{(K)}$, let $\{r_1, r_2, ..., r_K\}$ be the original index of $\{p_{(1)}, p_{(2)}, \ldots, p_{(K)}\}$ in the unsorted sequences $\{p_k, k = 1, \ldots, K\}$, so that $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_K}$.

Step 4. For a pre-specified level of p-value τ, denote r_T as the index of largest p-value smaller than τ, we accept all tests with index $r_T, r_{T+1}, ..., r_K$. And keep the T test with p-value smaller than τ into the proceeding steps, so that there are only T tests left with original index $r_1, ...r_T$.

Step 5. Perform randomization (e.g., bootstrap or permutation of group labels) to the

retained basis coefficients, and re-run test in Steps 2 using the randomized coefficients. Denote $\{p_k^*\}$ the resulting p-values; put $\{p_k^*\}$ in the same order as the original p-values, and denote $\{p^*_{(j)}, j = 1, \ldots.\}$. Note that $p^*_{(j)} = p^*_{r_j}, j \in \{1, ...T\}$.

Step 6. Repeat Step 4 for N times. Compute

$$q^*_{(j),l} = \min_{s \geq j} p^*_{(s)}, \quad l = 1, ..., N.$$

Note that $q^*_{(j),l}$ is an empirical version of $\min_{l \in \{j,...,T} P_{(l)}$.

Step 7. Calculate the frequency:

$$\widetilde{p}^{(N)}_{(j)} = \frac{1}{N} \sum_{l=1}^{N} I\{q^*_{(j),l} \leq p_{(j)}\}.$$

Note that this is an empirical approximation of $\Pr\left(\min_{l \in \{j,...,T\}} P_l \leq p_{(j)} \mid H_0^C\right)$ that is used in the definition of *free step-down adjusted p-value* in Westfall and Young [70, ch.2; page 66]. Here the capitalized P denote the random p-value, whose distribution is induced by the distribution of the data $\{y_{ij}(x)\}$.

Step 8. Enforce monotonicity using successive maximization:

$$\begin{aligned}
\widetilde{p}^{(N)}_{(1)} &= \widetilde{p}^{(N)}_{(1)} \\
\widetilde{p}^{(N)}_{(2)} &= \max(\widetilde{p}^{(N)}_{(1)}, \widetilde{p}^{(N)}_{(2)}) \\
&\vdots \\
\widetilde{p}^{(N)}_{(T)} &= \max(\widetilde{p}^{(N)}_{(T-1)}, \widetilde{p}^{(N)}_{(T)})
\end{aligned}$$

The reason for enforcing monotonicity was explained in Westfall and Young [70, ch.2; page 64].

A.2 Proof of Theorem 3.2

Based on the Assumptions 2, we will follow the proof of asymptotic optimality provided by Meinshausen et al. [42] to prove the optimaility of basis-space testing proposed described in Theorem 3.2. Before proving, we first introduce some additional notations:

- **Number of blocks** Simplify the number of blocks corresponding to K tests B_K by B.

- **Block-wise minimum p-value** Under the true null hypotheses, the minimum p-value of block b $(b = 1, \ldots, B)$ is defined as

$$p^{(b)}(W) = \min_{j \in A_b \cap S_0} p_j(W).$$

- **Cumulative distribution function of block-wise minimum p-value** Under the true null hypotheses, the cumulative distribution function of $p^{(b)}(W)$ $(b = 1, \ldots, B)$ is defined as

$$\pi_b(c) = P_K(p^{(b)}(W) \leq c).$$

Proof. Let $\alpha' \in (0, 1)$, $\delta' \in (0, \alpha')$. Let $\delta = \frac{\delta'}{2}$ and $\alpha = \alpha' - \delta'$. Then the expression

$$P_K\{\hat{c}_{K,n}(\alpha) \geq c_{K,n}(\alpha - \delta)\} \to 1 \text{ as } K \to \infty$$

can be written as

$$P_K(\hat{c}_{K,n}(\alpha - \delta + 2 \cdot \frac{\delta}{2}) \geq c_{K,n}(\alpha - \delta)) \to 1 \text{ as } K \to \infty.$$

Let $\tilde{\alpha} = \alpha - \delta$ and $\tilde{\delta} = \frac{\delta}{2}$. Then, replace $\tilde{\alpha}$ and $\tilde{\delta}$ by α and δ respectively. We have,

$$P_K(\hat{c}_{K,n}(\alpha + 2\delta) \geq c_{K,n}(\alpha)) \to 1 \text{ as } K \to \infty.$$

By definition,

$$\hat{c}_{K,n}(\alpha + 2\delta) = \max\{s \in \widehat{S_n} : P_K(\min_{j \in 1,\ldots,K} p_j(W) \leq s) \leq \alpha + 2\delta\}.$$

We thus have to show that

$$P_K(P^*(\min_{j=1,\ldots,K} \leq c_{K,n}(\alpha)) \leq \alpha + 2\delta) \to 1 \; as \; K \to \infty.$$

(i) First, we show in Lemma 1 (Appendix A.3) that there exists an $M < \infty$ s.t.,

$$P^*(\min_{j \in \{1,\ldots,K\}} p_j(W) \leq c_{K,n}(\alpha)) \leq P^*(\min_{j \in S_0} \leq c_{K,n}(\alpha)) + \delta$$

for all $K > M$ and for all W. This result is mainly due to the sparsity assumption (Assumption 2A.2).

(ii) Second, we show in Lemma 2 (Appendix A.4) that

$$P_K\{P^*(\min_{j \in S_0} p_j(W) \leq c_{K,n}(\alpha)) \leq \alpha + \delta\} \to 1 \; for \; K \to \infty.$$

If (i) - (ii) hold, we have

$$P_K[\{P^*(\min_{j \in S_0} p_j(W) \leq c_{K,n}(\alpha)) + \delta \leq \alpha + 2\delta\}] \leq P_K[\{P^*(\min_{j \in \{1,\ldots,K\}} p_j(W) \leq c_{K,n}(\alpha)) \leq \alpha + 2\delta\}].$$

By (ii), we can find $\tilde{M}$ s.t. when $K > \tilde{M}$, for $\forall \epsilon > 0, \forall \delta > 0$,

$$1 - P_K[\{P^*(\min_{j \in S_0} p_j(W) \leq c_{K,n}(\alpha)) + \delta \leq \alpha + 2\delta\}] \leq \epsilon.$$

Choose $\tilde{\tilde{M}} = \max\{M, \tilde{M}\}$. When $K \geq \tilde{\tilde{M}}$, we have

$$\begin{aligned} & 1 - P_K\{P^*(\min_{j \in \{1,\ldots,K\}} \leq c_{K,n}(\alpha)) \leq \alpha + 2\delta\} \\ \leq\ & 1 - P_K\{P^*(\min_{j \in S_0} p_j(W) \leq c_{K,n}(\alpha)) + \delta \leq \alpha + 2\delta\} \\ & \leq \epsilon. \end{aligned}$$

Thus, $P_K\{P^*(\min_{j \in \{1,\ldots,K\}} p_j(W) \leq c_{K,n}(\alpha)) \leq \alpha + 2\delta\} \to 1$, as $K \to \infty$.

Proof of (i) is in Lemma 1(Appendix A.3).

Proof of (ii) is in Lemma 2(Appendix A.4). □

A.3 Lemma 1

There exists an $M < \infty$ s.t.,

$$P^*(\min_{j\in\{1,\ldots,K\}} p_j(W) \leq c_{K,n}(\alpha)) \leq P^*(\min_{j\in S_0} \leq c_{K,n}(\alpha)) + \delta$$

for all $K > M$ and for all W.

Proof. We know that $c_{K,n} \in S_n$. For all $s \in S_n$ and all W, since

$$\begin{aligned}\{\min_{j\in\{1,\ldots,K\}} p_j(W) \leq s\} &= \{\min_{j\in S_0} p_j(W) \leq s\} \cup \{\min_{j\in S_0^C} p_j(W) \leq s\} \\ &= \{\min_{j\in S_0} p_j(W) \leq s\} \cup \{\bigcup_{j\in S_0^C} p_j(W) \leq s\},\end{aligned}$$

Thus,

$$P^*\{\min_{j\in\{1,\ldots,K\}} p_j(W) \leq s\} \leq P^*\{\min_{j\in S_0} p_j(W) \leq s\} + \sum_{j\in S_0^C} P^*\{p_j(W) \leq s\}.$$

To show this lemma, we just need to show that $\exists M < \infty$, s.t., for all $m > M$, and all W,

$$\sum_{j\in S_0^C} P^*\{p_j(W) \leq c_{K,n}(\alpha)\} \leq \delta.$$

By Assumption 2B.2, $\exists r < \infty$, s.t.,

$$\begin{aligned}\sum_{j\in S_0^C} P^*\{p_j(W) \leq c_{K,n}(\alpha)\} &\leq |S_0^C| c_{K,n}(\alpha) r \\ &\leq r \cdot \frac{|S_0^C|}{B} \cdot B c_{K,n}(\alpha).\end{aligned}$$

By Assumption 2A.2, $\frac{|S_0^C|}{B_K} \to 0$ as $K \to \infty$.

Now look at $B \cdot c_{K,n}(\alpha)$. In Lemma 3 (Appendix A.5), we have shown that

$$B \cdot c_{K,n}(\alpha) \leq -log(1-\alpha).$$

So we can choose a large M', so that

$$.r \cdot \frac{|S_0^C|}{B} B c_{K,n}(\alpha) \leq \delta, \ for \ \forall K > M'.$$

□

A.4 Lemma 2

$$P_K\{P^*(\min_{j\in S_0} p_j(W) \le c_{K,n}(\alpha)) \le \alpha+\delta\} \to 1 \ for \ K\to\infty.$$

Proof. Let $\epsilon > 0$. This lemma is equivalent that $\exists M < \infty$, s.t.,

$$P_K\{P^*(\min_{j\in S_0} p_j(W) > c_{K,n}(\alpha)) < 1-\alpha-\delta\} < \epsilon, \ for \ \forall m > M.$$

By definition,

$$P^*(\min_{j\in S_0} p_j(W) > c_{K,n}(\alpha)) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \mathbb{1}\{\min_{j\in S_0} p_j(gW) > c_{K,n}(\alpha)\} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} R(g,W).$$

Also, by Assumption 2B.1, we have $E_K(\frac{1}{|\mathcal{G}|R(g,W}) = E_{K,G}R(G,W) = E_K R(1,W)$, and $E_K R(1,W) = P_K(\min_{j\in S_0 p_j(W} > c_{K,n}(\alpha)) > 1-\alpha$. Notice that,

$$\begin{aligned}
& P_K\{P^*(\min_{j\in S_0} p_j(W) > c_{K,n}(\alpha)) < 1-\alpha-\delta\} \\
= \ & P_K\{\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} R(g,W) < 1-\alpha-\delta\} \\
= \ & P_K\{\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} R(g,W) - E_K R(1,W) < 1-\alpha-\delta-E_K R(1,W)\} \\
\le \ & P_K\{\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} R(g,W) - E_K R(1,W) < -\delta\} \\
\le \ & P_K\{|\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} R(g,W) - E_K R(1,W)| > \delta\} \\
\le \ & \frac{1}{\delta^2} Var(\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} R(g,W)).
\end{aligned}$$

To show the above smaller than ϵ, we just need to show:

$$Var(\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} R(g,W)) = o(1),\ as\ K\to\infty.$$

Also,$Var(\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} R(g,W)) = \frac{1}{|\mathcal{G}|^2}\sum_{g,g'\in\mathcal{G}} Cov_K(R(g,W),R(g',W))$. Note that, it could be $g=g'$.

Let G and G' be two random permutations, drawn independently and uniformly from $\mathcal{G}$. Then,

$$Cov_{K,G,G'}(R(G,W),R(G',W)) = \frac{1}{|\mathcal{G}|^2}\sum_{g,g'\in\mathcal{G}} Cov_K(R(g,W),R(g',W)).$$

Thus, it is equivalent to showing that

$$Cov_{K,G,G'}(R(G,W),R(G',W)) = o(1)\ for\ K\to\infty.$$

Define $R_b(g,W) = \mathbb{1}\{p^{(b)}(g,W) > c_{K,n}(\alpha)\}$, where $p^{(b)}(g,W) = \min_{j\in A_b\cap S_0} p_j(W)$. So,

$$R(g,W) = \prod_{b=1}^{B} R_b(g,W).$$

We then need to prove that

$$E_{K,G,G'}(\prod_{b=1}^{B} R_b(G,W)R_b(G',W)) - (E_{K,G}\{\prod_{b=1}^{B} R_b(G,W)\})^2 = o(1) \tag{A.1}$$

By assumption 2A.1, the left-hand side can be written as

$$\prod_{b=1}^{B} E_{K,G,G'}(R_b(G,W)R_b(G',W)) - \prod_{b=1}^{B}(E_{K,G}\{R_b(G,W)\})^2.$$

Note that both $E_{K,G,G'}(R_b(G,W)R_b(G',W))$ and $(E_{K,G}\{R_b(G,W)\})^2$ are bounded between 0 and 1. The following inequality holds:

$$|\prod_{j=1}^{B} a_j - \prod_{j=1}^{B} b_j| = |\sum_{j=1}^{B}\{(a_j - b_j)(\prod_{k<j} b_k)(\prod_{k>j} a_j)\}| \leq \sum_{j=1}^{B} |a_j - b_j|.$$

In order to show Equation A.1, it is sufficient to show that

$$\max_{b=1,\ldots,B} |E_{K,G,G'}\{R_b(G,W)R_b(G',W)\} - (E_{K,G}\{R_b(G,W)\})^2| = o(B) \tag{A.2}$$

as $K \to \infty$.

Conditional on W, $R_b(G,W)$, $R_b(G',W)$ are i.i.d. random variables from $Bernoulli(\mu_b(W)$. Since the distribution of G is discrete counting measure, $\mu_b(W)$ is the proportion of $R_b(g,W) = 1$, i.e.,

$$\mu_b(W) = P_{K,G}\{p^{(b)}(GW) > c_{K,n}(\alpha)|W\} = P^*\{p^{(b)}(W) > c_{K,n}(\alpha)|W\}.$$

Thus, $\mu_b(W)$ is a random proportion. Its distribution can be denoted by F_b, $F_b : [0,1] \to [0,1]$. By Lemma 4 (Appendix A.6), the support of F_b is contained in $[1 - log(\frac{1}{1-\alpha})\alpha r^2 m_B \cdot B^{-1}, 1]$. By Lemma 5 (Appendix A.7), it follows that

$$0 \leq E_{K,G,G'}\{R_b(G,W)R_b(G',W)\} - (E_{K,G}\{R_b(G,W)\})^2 \leq (log(\frac{1}{1-\alpha})\alpha r^2 m_B \cdot B^{-1})^2.$$

By Assumption 2A.3, $m_B = o(\sqrt{B})$, so Equation A.2 holds. □

A.5 Lemma 3

Under assumptions 2A.1 and 2B.3, we can show that

$$Bc_{K,n}(\alpha) \leq \sum_{b=1}^{B} \pi_b(c_{K,n}(\alpha)) \leq log(\frac{1}{1-\alpha}), \tag{A.3}$$

where $\pi_b(c_{K,n}(\alpha)) := P_K\{p^{(b)}(W) \leq c_{K,n}(\alpha)\}$.

Proof. Let $b \in \{1, \ldots, B\}$ and $j_b \in S_0 \cap A_b$. Then,

$$\begin{aligned} \pi_b(c_{K,n}(\alpha)) &= P_K\{p^{(b)}(W) \leq c_{K,n}(\alpha)\} \\ &= 1 - P_K\{p^{(b)}(W) > c_{K,n}(\alpha)\} \\ &\geq P_K\{p_{j_b}(W) \leq c_{K,n}(\alpha)\}. \end{aligned}$$

Thus,

$$\pi_b(c_{K,n}(\alpha)) \geq P_K\{p_{j_b}(W) \leq c_{K,n}(\alpha)\} \overset{2B.3}{=} c_{K,n}(\alpha).$$

$$\implies \sum_{b=1}^{B} \pi_b(c_{K,n}(\alpha)) \geq Bc_{K,n}(\alpha).$$

To prove the second inequality of this lemma, note that:

$$\begin{aligned} & P_K\{\min_{j \in S_0} p_j(W) \leq c_{K,n}(\alpha)\} \leq \alpha \\ \implies & 1 - P_K\{\min_{j \in S_0} p_j(W) > c_{K,n}(\alpha)\} \leq \alpha \\ \implies & 1 - \prod_{b=1}^{B} P_K\{p^{(b)}(W) > c_{K,n}(\alpha)\} \leq \alpha \\ \implies & 1 - \prod_{b=1}^{B} (1 - \pi_b(c_{K,n}(\alpha))) \leq \alpha \end{aligned} \tag{A.4}$$

If we consider to maximize $\sum_{b=1}^{B} a_b$ under the constraint $1 - \prod_{b=1}^{B}(1 - a_b) \leq \alpha$. Using Lagrange multiplier, we see the maximum of $\sum_b a_b$ is obtained when $a_1 = \ldots a_B$.
Thus, $\sum_{b=1}^{B} \pi_b(c_{K,n}(\alpha))$ is maximized under constraint A.4 if

$$\pi_1(c_{K,n}(\alpha)) = \ldots = \pi_B(c_{K,n}(\alpha))$$

. This implies

$$\begin{aligned} & 1 - \prod_{b=1}^{B}(1 - \pi_b(c_{K,n}(\alpha))) \leq \alpha \\ \implies \quad & \pi_b(c_{K,n}(\alpha)) \leq 1 - (1-\alpha)^{1/B} \\ \implies \quad & B\pi_b(c_{K,n}(\alpha)) \leq B - B(1-\alpha)^{1/B}. \end{aligned}$$

$B - B(1-\alpha)^{1/B}$ is bounded above by $-log(1-\alpha)$ for all B. Thus,

$$\sum_{b=1}^{B} \pi_b(c_{K,n}(\alpha)) \leq -log(1-\alpha).$$

□

A.6 Lemma 4

Suppose that $\mu_b(W) = P^*\{p^{(b)}(W) > c_{K,n}(\alpha)\}$. Let F_b be the distribution of $\mu_b(W)$. Then,

$$support(F_b) \subseteq [1 - log(\frac{1}{1-\alpha})\alpha r^2 m_B B^{-1}, 1].$$

Proof. By Assumption 2B.2,

$$\begin{aligned}
1 - \mu_b(W) &= P^*\{p^{(b)}(W) \leq c_{K,n}(\alpha)\} \\
&= P^*\{\min_{j \in A_b \cap S_0} p_j(W) \leq c_{K,n}(\alpha)\} \\
&= P^*\{\bigcup_{j \in A_b \cap S_0} p_j(W) \leq c_{K,n}(\alpha)\} \\
&= \sum_{j \in A_b \cap S_0} P^*\{p_j(W) \leq c_{K,n}(\alpha)\} \\
&= |A_b| \cdot r \cdot c_{K,n}(\alpha)
\end{aligned}$$

This implies that $1 - |A_b| r c_{K,n}(\alpha) \leq \mu_b(W) \leq 1$.

So, we just need to show that $1 - |A_b| r c_{K,n}(\alpha) \leq \mu_b(W) \geq 1 - log(\frac{1}{1-\alpha})\alpha r^2 m_B \cdot B^{-1}$. i.e.,

$$|A_b| r c_{K,n}(\alpha) \leq \mu_b(W) \leq log(\frac{1}{1-\alpha})\alpha r^2 m_B \cdot B^{-1}.$$

By assumption 2A.3, we know that $|A_b| \leq m_B$. This leaves to prove that

$$c_{K,n}(\alpha) \leq log(\frac{1}{1-\alpha})\alpha r B^{-1}. \tag{A.5}$$

To show Equation A.5, we first show that

$$1 - \alpha \leq P_K\{\min_{j \in S_0} p_j(W) > c_{K,n}(\alpha)\} \leq (1 - c_{K,n}(\alpha)/r)^B \tag{A.6}$$

To show the second inequality, we notice that assumption 2A.2 implies

$$P_K\{\min_{j\in S_0} p_j(W) > c_{K,n}(\alpha)\} = \prod_{b=1}^{B} P_K\{p^{(b)}(W) > c_{K,n}(\alpha)\},$$

and assumption 2B.1 implies

$$P_K\{p^{(b)}(W) > c_{K,n}(\alpha)\} = P_{K,G}\{p^{(b)}(GW) > c_{K,n}(\alpha)\} = E_K\{P_{K,G}\{p^{(b)}(GW) > c_{K,n}(\alpha)|W\}\}. \tag{A.7}$$

Then, we have:

$$\begin{aligned} P_{K,G}\{p^{(b)}(GW) > c_{K,n}(\alpha)|W\} &= P^*\{p^{(b)}(W) > c_{K,n}(\alpha)\} \\ &\leq P^*\{p_{j_b}(W) > c_{K,n}(\alpha)\} \\ &= P^*\{p_{j_b}(W) \leq c_{K,n}(\alpha)\} \\ &\leq 1 - c_{K,n}(\alpha)/r \end{aligned} \tag{A.8}$$

Here $j_b \in S_0 \cap A_b$. Since the right hand side of Inequality A.8 does not depend on W, the same bound holds for Inequality A.7. Thus, we have

$$P_K\{\min_{j\in S_0} p_j(W) > c_{K,n}(\alpha)\} \leq (1 - c_{K,n}(\alpha)/r)^B.$$

1.e., Inequality A.6 holds. Solve $c_{K,n}(\alpha)$ from A.6, we get:

$$c_{K,n}(\alpha)\} \leq r\{1 - (1-\alpha)^{1/B}\}.$$

Since $B(1-(1-\alpha)^{1/B}) \leq -log(1-\alpha)$ for all B, it follows that

$$1 - (1-\alpha)^{1/B} \leq -log(1-\alpha) \cdot B^{-1},$$

which implies that

$$c_{K,n}(\alpha) \leq -log(1-\alpha) \cdot B^{-1} \cdot r$$

. Thus, Inequality A.5 holds. □

A.7 Lemma 5

Suppose U is a real-valued random variable with support $[a, b] \subset [0, 1]$ and we have

$$X_1, X_2 | U = u \overset{i.i.d}{\sim} Bernoulli(u).$$

Then,

$$0 \leq E[X_1 X_2] - E[X_1]E[X_2] \leq (b - a)^2.$$

Proof. By assumption,

$$\begin{aligned}
&E[X_1|U] = U,\ E[X_2|U] = U,\\
&E[X_1 X_2] = E[E[X_1 X_2|U]] = E[E[X_1|U]E[X_2|U]] = E[U^2],\\
&E[X_1] = E[E[X_1|U]] = E[U],\ E[X_2] = E[U].
\end{aligned}$$

Thus, $E[X_1 X_2] - E[X_1]E[X_2] = E[U^2] - (E[U])^2 = Var(U)$. Since the support of U is $[a, b] \subset [0, 1]$, we have

$$Var(U) = E(U - E[U])^2 = \int_a^b (U - E[U])^2 dF(u) \leq (b - a)^2 \int_a^b dF(u).$$

□