2019

Methods for Analyzing High Dimensional Data with Applications to the Wearable and Microbiome Data Analysis

Quy Xuan Cao

Methods for Analyzing High Dimensional Data with Applications to the Wearable and Microbiome Data Analysis

By

Quy Xuan Cao

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in Mathematics

The University of Montana
Missoula, MT

November 2019

Approved by:

Ashby Kinch, Associate Dean of Graduate School
Graduate School

Dr. Ekaterina Smirnova, Chair
Mathematical Sciences

Dr. Leonid Kalachev
Mathematical Sciences

Dr. Jonathan Graham
Mathematical Sciences

Dr. Johnathan Bardsley
Mathematical Sciences

Dr. Nathan Insel
Department of Psychology

Methods for Analyzing High Dimensional Data
with Applications to the Accelerometry and Microbiome Data

Quy Xuan Cao

University of Montana, 2019

ABSTRACT

Modern studies in medicine, epidemiology, pharmacy and other fields generate high dimensional data. We developed statistical analysis methods for two types of such data: activity and microbiome data. Specifically, reliable measures of the frequency, duration and intensity of physical activity provided by wearable technology were used in the analysis of activity data. Accelerometry-derived measures of physical activity were compared with established predictors of 5-year all-cause mortality in older adults, aged between 50 and 85 years from the 2003- 2006 National Health and Nutritional Examination Survey, in terms of individual, relative, and combined predictive performance. A total of 33 predictors of 5-year all-cause mortality, including 20 measures of objective physical activity, were compared using single-predictor and multiple logistic regression. The results show that objective accelerometry-derived physical activity measures outperform traditional predictors of 5-year mortality in single predictor models, and offer some improvement in multiple predictor models beyond what age and other traditional predictors provide. This highlights the importance of wearable technology for providing reproducible, unbiased, and prognostic biomarkers of health. In microbiome data, we concentrated on pre-processing steps, where both the sparsity of counts and the large number of observed taxa were considered. The current approach is to remove taxa that appear in small counts in a few samples, which is known as filtering. We

present the package **PERFect** which performs a permutation filtering approach designed to address two problems in microbiome data processing: (1) define and quantify loss due to filtering by implementing thresholds; and (2) introduce and evaluate a permutation test for filtering loss to provide a measure of excessive filtering. The package employs an unbalanced binary search algorithm that greatly reduces computational time for these permutations. The effectiveness of the proposed approach on downstream microbiome data analysis is illustrated on two microbiome quality control datasets. Our filtering method reduces: (1) the magnitude of differences in alpha diversity for samples containing the same bacteria processed at different labs and (2) the dissimilarity between samples (beta diversity) that contain the same microbiome potentially alleviating technical variability.

Keywords: High Dimensional, Accelerometry, Physical Activity, Physical Performance, Exercise, Longevity, Microbiome, Filtering, Permutation Test, Binary Search, Skew-normal Distribution, Quality Control.

Acknowledgments

The completion of this dissertation would not have been possible without the support of those around me. First, I would like to express my deepest gratitude to my advisor, Dr. Ekaterina Smirnova, for her encouragement and essential guidance at every step of the dissertation. Her tremendous help on my research as well as her continued support has led me to the right direction and finish my dissertation. I would also like to extend my appreciation to my committee members: Dr. Jonathan Graham, Dr. Leonid Kalachev, Dr. Johnathan Bardsley, and Dr. Nathan Insel for spending their invaluable time to evaluate my dissertation and provide me with valuable feedback. My sincere appreciation to all the authors who published their papers. They help me enormously in understanding the subject and writing my dissertation.

Finally, I would also like to extend gratitude to my friends and family. Their unfathomable supports during the process helped me achieve my goals.

Contents

List of Figures

List of Tables

List of Notations

AIC	Akaike's Information Criterion
ASTP	Active to Sedentary Transition Probability
AUC	Area Under the Curve
BMI	Body Mass Index
CDC	Centers for Disease Control
CHD	Coronary Heart Disease
CHF	Congestive Heart Failure
CI	Confidence Interval
DFL	Difference in Filtering Loss
EPIC	Efficient Parsimony Information Criterion
FDA	Functional Data Analysis
FL	Filtering Loss
fMRI	functional Magnetic Resonance Imaging
fPCA	functional Principal Component Analysis
HMP	Human Microbiome Project
ICL	Imperial College London
LEfSe	Linear Discriminant Analysis Effect Size
MBQC	MicroBiome Quality Control
MDS	Multidimensional Scaling

MISE	Mean Integrated Squared Error
m_{i1}	Average scores for the first principal component
MSE	Mean Squared Error
mV	Millivolts
MVPA	Moderate/Vigorous Physical Activity
NB	Negative Binomial
NRI	Net Reclassification Index
NDI	National Death Index
NGS	Next Generation Sequencing
NHANES	National Health and Nutrition Examination Survey
OR	Odd Ratio
OTU	Operational Taxonomic Unit
PA	Physical Activity
PC	Princial Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
ROC	Receiver Operating Characteristic
SATP	Sedentary to Active Transition Probability
SBP	Systolic Blood Pressure
SIBO	Small Intestine Bacteria Overgrowth
s_{i6}	Standard deviation of the sixth principal component
SN	Skew-Normal
TAC	Total Activity Count
TLAC	Total Log Activity Count
UB	University of Birmingham

WTSI	Wellcome Trust Sanger Institute
ZIG	Zero-Inflated Gaussian
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson

Chapter 1

Introduction

1.1 High dimensional data

The rise of technological developments has shifted research towards heavier use of computational tools. It began in the late 1960s, when academics started using statistical software like SPSS to perform complex computations instead of manual calculations [SPSS, 2018]. This addition of technology to the research process has reduced the potential for human error and increased computational speed. Several decades later, in the 2000s, the spectacular evolution of data acquisition technologies and computing facilities started changing the way researchers collect and analyze data [Johnstone and Titterington, 2009]. From the classical scenario of 'small p, large n' (p is the number of variables and n is the number of observations), modern data have become 'large p, small n' or 'large p, large n', which is now referred to as high dimensional data. This introduced new complex analyses that include image analysis, genomics research, document classification, and so on. Hence, the need for new data analysis methods to provide computational efficiency and practical results

have increased in response to these changes.

Studies using high dimensional data have shown many significant breakthroughs in medicine, epidemiology, pharmacy and other fields. One example of high dimensional data is the activity measure using accelerometry, which has received growing attention in recent years as shown in Figure 1.1. For example, it is extremely difficult for field biologists to track wild animals' activities; thus their attempts to quantify behavior to model ecological processes may be inaccurate due to the lack of observing important behavioral events. Using acceleration sensors, researchers can now measure the change in velocity of a body over time as well as quantify fine-scale movements and body postures without issues of visibility, observer bias, or the scale of space use [Brown et al., 2013]. Moreover, the technology and application of current accelerometer-based devices in human physical activity (PA) research allow the capture and storage of large volumes of raw acceleration signal data, which provide opportunities to characterize and improve physical activity behavioral patterns [Troiano et al., 2014]. Functional magnetic resonance imaging (fMRI) is another area where high dimensional data arise. Studies in fMRI analyze functional brain networks to better understand how brain regions interact, how this depends upon experimental conditions and behavioral measures and how anomalies (disease) can be recognized [Solo et al., 2018]. Lastly, microbiome data are known to be high dimensional due to the number of samples and bacteria identified as a result of the sequencing process. The analyses of the associations between the human microbiome and health aim to understand the host-microbiome interactions and integrate them with other 'omics' datasets to enhance precision medicine [Petrosino, 2018].

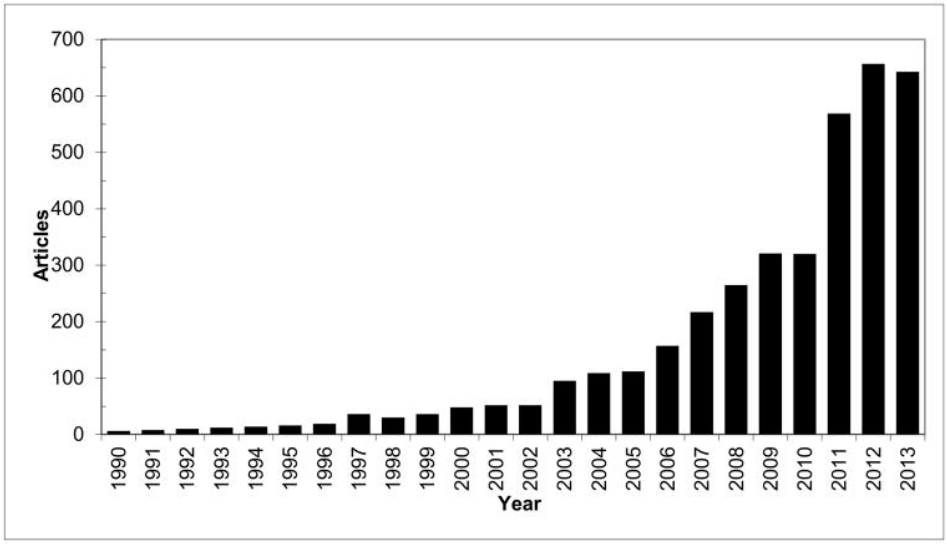


Figure 1.1: Publications by year with search terms 'exercise or physical activity' and 'accelerometry'. Source: [Troiano et al., 2014].

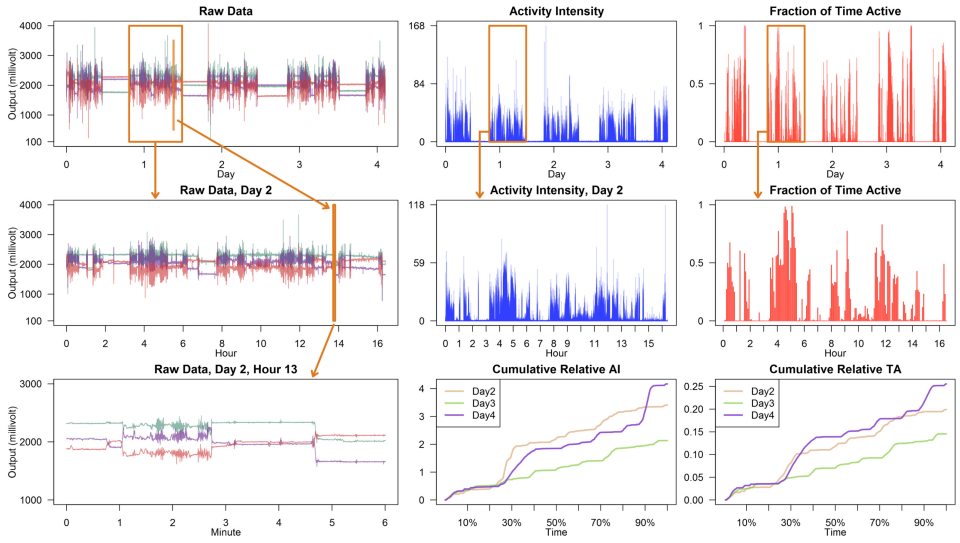


Figure 1.2: Accelerometry data for one subject followed over 5 days. Source: [Smirnova et al., 2018b].

1.2 Accelerometry data

In this dissertation, two types of high dimensional data are considered for analysis: accelerometry data and microbiome data. Figure 1.2 displays accelerometry data collected at a frequency of 10 Hz for five days from a sensor placed on the hip of a person [Bai et al., 2012]. The top left panel shows data measured along three orthogonal axes (up-down, left-right, backward-forward in the device frame of reference) for one subject, whose data consist of approximately 13 million observations. The five days of long periods of higher amplitude signals are separated by four nights characterized by low amplitude signals. To get a closer view, the box in the top-left panel in Figure 1.2 which identifies day 2 of the data is zoomed in as shown in the left-middle panel. A vertical line indicates a period of six minutes during day 2, which is further zoomed in and shown in the left-bottom panel. As one looks at finer resolutions of the data, more patterns can be identified and possibly used. These raw data are expressed in millivolts (mV), though most devices output raw data in Earth gravitational units ($g = 9.81 m/s^2$). Working directly with raw data could be quite daunting and, in practice, data are often summarized as activity counts (or steps) per minute as in Figure 1.3, resulting in a matrix of $n \times 1440$, where n is the number of samples and 1440 represents the number of minutes per day. The middle-top panel in Figure 1.2 provides such a summary measure at the minute level, while the middle-center panel displays the same measure for day 2. While informative, overlaying such visualizations in the same panel will lead to over-plotting and loss of information when comparing different days or subjects or when displaying an entire cohort. Instead, the middle-bottom plot shows the cumulative measure of activity up to a partic-

ular time of the day. This panel contains exactly the same information as the middle-center panel, but allows for joint plotting of multiple days and subjects. The right panels display similar information, although they focus on the proportion of time active per minute instead of activity intensity during that minute. The proportion of time active is obtained by calculating the activity intensity at the second level, applying a threshold on activity intensity that indicates active/inactive, and then computing the proportion of active seconds within that minute [Smirnova et al., 2018b] [Karas et al., 2019].

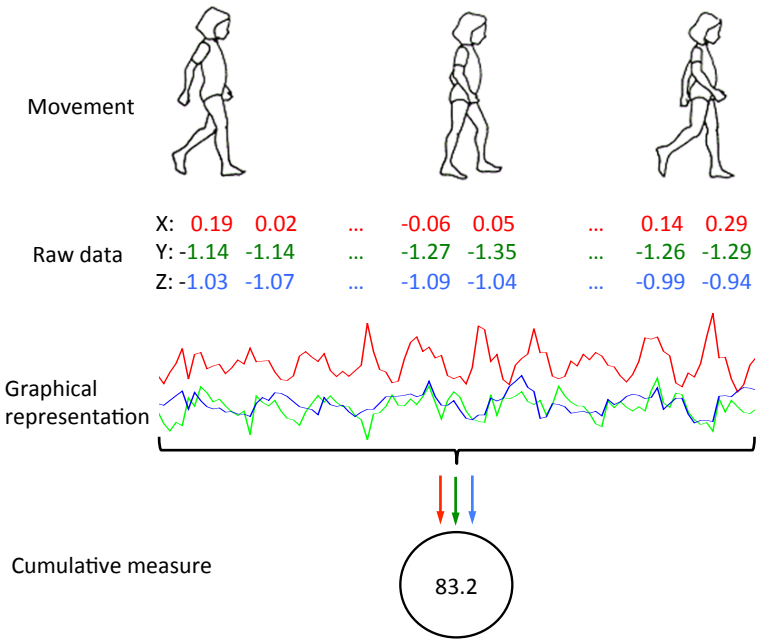


Figure 1.3: Raw data summarized as activity counts

Traditionally, the physical activity data resulting from accelerometry measures are analyzed using methods in functional data analysis (FDA), which

deals with data that are in the form of continuous functions [Augustin et al., 2017], [Smirnova et al., 2019] and [Leroux et al., 2019]. Here, each function is typically observed at a finite number of points, and in the case of accelerometry data summarized at the minute level, these functional data are observed throughout 1440 points (minutes) in a day. Key aspects of FDA include the choice of smoothing technique, dimension reduction, adjustment for clustering and functional linear modeling [Finch, 2013]. The first step in any FDA is smoothing, which represents raw discrete data points as a smooth function that emphasizes patterns in the data by minimizing noise due to observational errors. In particular, the use of B-spline basis functions is one of the most popular smoothing techniques [Aguilera and Aguilera-Morillo, 2013]. As for data reduction, functional principal components analysis (fCPA) is a popular multivariate analysis technique for extracting information from multiple variables by reducing the dimensions of a dataset while preserving as much of the total variation as possible [Croux and Ruiz-Gazen, 2005]. As fPCA results in dimension reduction, fPCA vector scores can be used for clustering different functions/components using standard clustering methods [Finch, 2013]. In the accelerometry data context, clustering helps to identify representative curve patterns and individuals with similar activity patterns. An interesting application of FDA involves the construction of functional linear models that describe the relationship between a response and explanatory variables [Usset et al., 2016]. Here, functions could be used as the response variable, the predictors or both. In **R**, the package **fda** [Ramsay et al., 2018] and **refund** [Goldsmith et al., 2018] provide various statistical tools for functional data analysis and are freely available for researchers to use.

1.3 Microbiome data

Microbiome data are another type of high dimensional data that will be discussed in this dissertation. To generate a microbiome dataset, the first step is to collect samples which could be taken at various body sites [Belizario and Napolitano, 2015], as shown in Figure 1.4. These samples are sequenced using the next generation sequencing (NGS) of the 16S ribosomal RNA genes technology to generate DNA fragments, which are then grouped into similar microbial organisms called taxa [Sanschagrin and Yergeau, 2014]. The resulting dataset, which has samples in the rows and taxa in the columns, is a large sparse matrix as many rare taxa are identified. In current microbiome studies, the goal is to understand mechanisms of host genetic and environmental factors that shape the microbiome. For example, in 2008, a multi-institutional collaboration called the Human Microbiome Project (HMP) was established to generate resources that facilitate characterization of the human microbiota and further our understanding of how the microbiome impacts human health and disease [Turnbaugh et al., 2007]. Specifically, this project aimed to develop a reference set of 3,000 isolate microbial genome sequences, understand the 'core' microbiome at five regions in the body (nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract), determine the relationship between disease and changes in the human microbiome and develop new tools and technologies for computational analysis [HMP, 2008]. However, it is difficult to reproduce studies across labs because variation in measurements between laboratories has not been systematically assessed. The Microbiome Quality Control (MBQC) project was therefore initiated to identify sources of variation in microbiome studies, to quantify their magnitudes,

and to assess the design and utility of different positive and negative control strategies [Sinha et al., 2015].

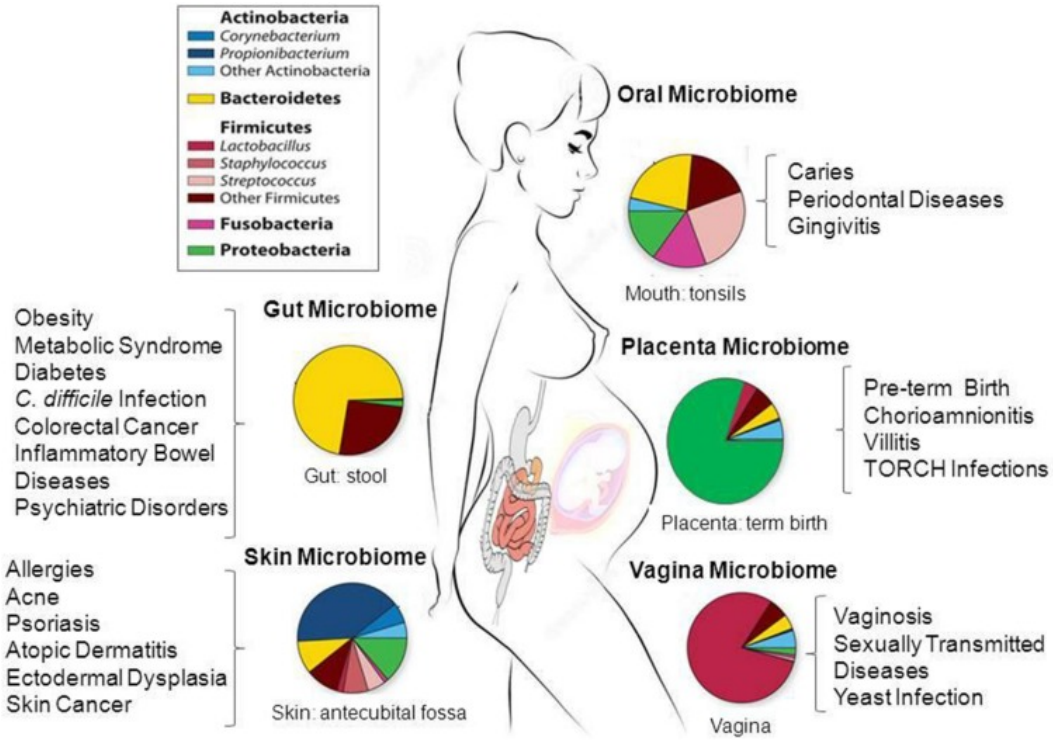


Figure 1.4: Prevalence and abundance of microbial taxa inhabiting healthy human body sites. Source: [Belizario and Napolitano, 2015]

Dynamic interactions exist among environment, microbiome and host. For microbiome studies, the focus is to test the association between the microbiome and the host, specifically whether the composition of the microbiome or 'dysbiotic' microbiome is linked to the health or disease of the host [Xia and Sun, 2017]. For example, in small intestine bacteria overgrowth (SIBO) research, dysbiosis is associated with the overgrowth of pathogenic bacteria in the small intestine, causing pain and diarrhea and leading to malnutrition [Leite et al., 2019]. It is also of interest to test whether the microbiome is associated with environmental covariates or whether there is

an effect of intervention of a specific microbiome composition on health and disease [Chen et al., 2012]. Examples include testing whether dietary interventions shape gut microbiota [Albenberg et al., 2012] and understanding the impact of a probiotic intervention on the composition of the human microbiota [Lahti et al., 2013]. However, when analyzing microbiome data, taxa counts are often overdispersed and have many zeros as displayed in Figure 1.5. In order to fit microbiome count data with overdispersion and excess zeros, the negative binomial (NB) [Zhang et al., 2018] and zero inflated models such as the Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial (ZINB) and Zero-inflated Gaussian (ZIG) mixture model [Paulson et al., 2017] were chosen for modeling the excess zeros and testing differential abundance taxa between groups.

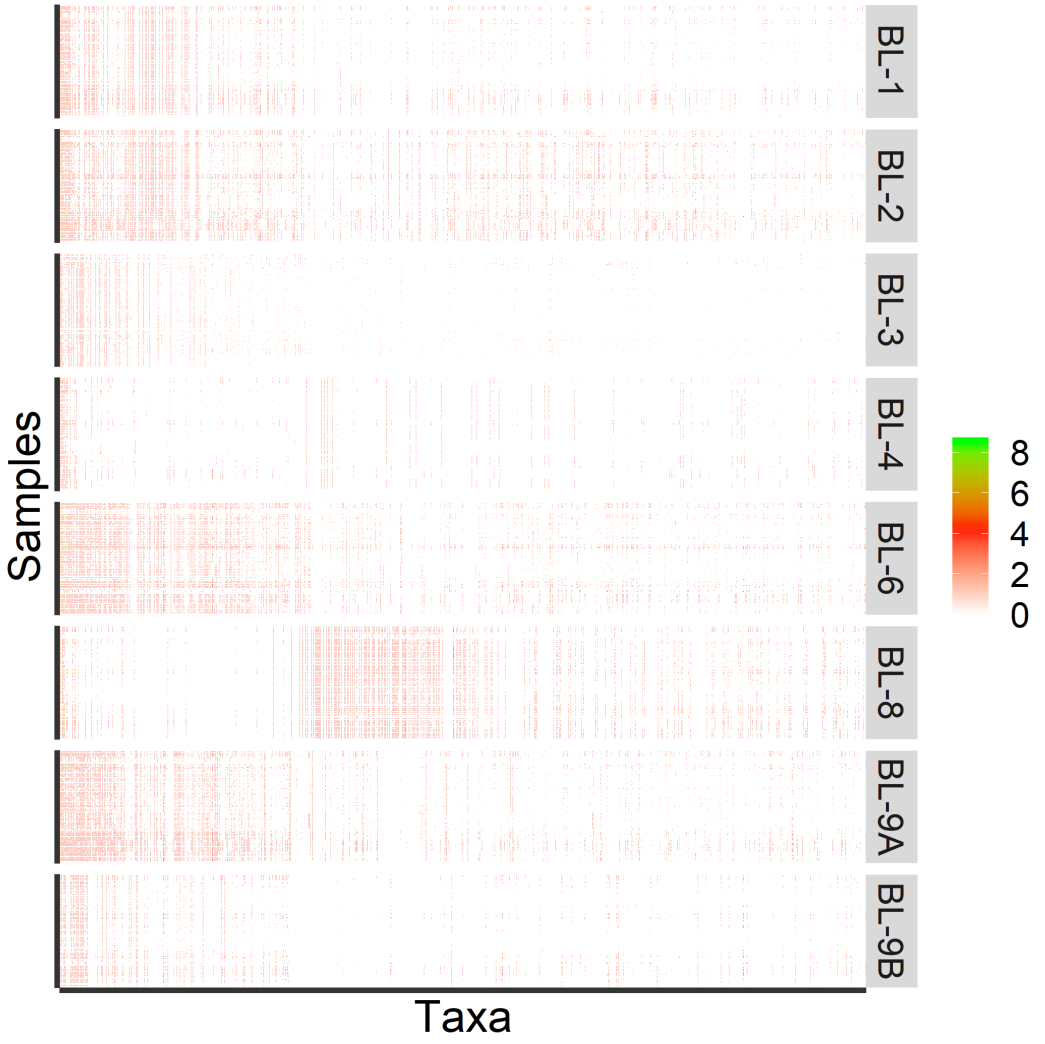


Figure 1.5: Heatmap of 1016 observed taxa on the log-scale, with taxa on the x-axis arranged in decreasing abundance order and samples on the y-axis arranged by processing institutes. Source: [Sinha et al., 2015]

1.4 Research contributions

The research problems studied as part of this dissertation, include the analysis of the National Health and Nutrition Examination Survey (NHANES)

accelerometry data as well as the development of methodology for the microbiome quality control problem. Specifically, the accelerometry data studies have mainly focused on developing interpretable metrics for summarizing raw tri-axial accelerometry data [Bai et al., 2013] [Bai et al., 2016], deriving appropriate measures for physical activity [Varma et al., 2018] and re-evaluating the effect of age on physical activity over a lifespan [Varma et al., 2017]. However, development of mortality predictive models using accelerometry data and the influence of physical activity while considering traditional predictors such as age and body mass index simultaneously remained open. Hence, for my research, I explored the associations between participants' physical activity, demographic and health-related characteristics and 5-year all-cause mortality in NHANES data, performed single-predictor logistic regression to identify the ranking of the most predictive predictors and their relative effects on mortality, and compared derived measures of physical activity to established predictors of 5-year all-cause mortality. With the assistance of my collaborators from Johns Hopkins University in deriving information criterion for complex survey design model, I was able to build a multiple logistic regression model using forward selection. Our results led to two publications and were featured in the recent press release of the Johns Hopkins University [JHU, 2019]. Moreover, I contributed to the **R** package **rnhanesdata** (available on github) which organizes and helps with the analysis of the Activity Data in NHANES [Leroux et al., 2019]. In this package, I helped with automating the extraction of data from the Centers for Disease Control and Prevention website, merging multiple files into one final file and cleaning the data based on our exclusion criteria. A detailed vignette that describes the data processing steps and analysis is publicly available within the package to guide researchers who plan to

use this package and replicate our findings.

For microbiome data, the microbiome quality control problem needs to be addressed prior to data analysis. Recent microbiome quality control studies show that the majority of rare taxa are caused by contamination and/or sequencing errors [Sinha et al., 2015]. The most common approach to address this problem is to filter spurious taxa from the data, and one of the most widely used techniques for filtering in microbiome studies is to select taxa that have a number of counts above $m = 0$ in at least n samples. [Davis et al., 2018] introduced the **decontam** R package that identifies contaminants using DNA concentration information which might not be always available. [Smirnova et al., 2018a] introduced a filtering test, PERFect, by filtering out taxa with insignificant contribution to the total covariance. However, the earlier software implementation was computationally intensive due to the complex permutation filtering algorithm. Hence, using the idea of an unbalanced binary search algorithm, I developed a fast implementation of this algorithm that optimally finds the set of taxa to be removed without building the permutation distribution and computing the p-values for all taxa [Morin, 2013]. The proposed approach successfully reduces the algorithm run time by almost four times. I also developed the **R** package **PERFect** which was published in Bioconductor, a free and open development software project for the analysis and comprehension of genomic data [PERFect, 2019]. The reference manual for this package can be found in the appendix of the dissertation. I then evaluated the effect of filtering on two major exploratory analyses used in microbiome research: alpha and beta diversity. The methods were applied to two data sets, namely the MicroBiome Quality Control (MBQC) project from [Sinha et al., 2015] and the laboratory contamination

dataset from [Salter et al., 2014]. Results show that the filtering methods reduce the magnitude of differences in alpha diversity for samples containing same bacteria processed at different labs. Filtering further reduces dissimilarity between samples (beta diversity) that contain the same microbiome and potentially alleviates technical variability. Results of this research are currently being prepared for publication.

The rest of the dissertation is organized as follows. I introduce the necessary background for functional data in Chapter 2. I show in Chapter 3 the application of functional data analysis in the NHANES data. The microbiome data and the filtering method PERFect are described in Chapter 4. Concluding remarks follow in Chapter 5.

Chapter 2

Functional Data Analysis

2.1 What is Functional Data?

2.1.1 Introduction

Functional data analysis corresponds to analysis of information on continuous functions (or curves), typically observed at a finite number of points. The primary interest is to study the behavior of such data, and their relationship to other quantities. For example, Figure 2.1 displays average monthly temperature and precipitation data at 35 different locations in Canada averaged over 1960 to 1994 [Ramsay and Silverman, 2006]; each smoothed curve can be considered as a function of temperature and precipitation of each location over time. Another example from [Ramsay and Silverman, 2006] is the Berkeley growth study data, displayed in Figure 2.2, in which the heights of boys and girls were recorded from age 1 to 18. Here, heights of boys and girls can be treated as smoothed functions over the 18 recorded ages.

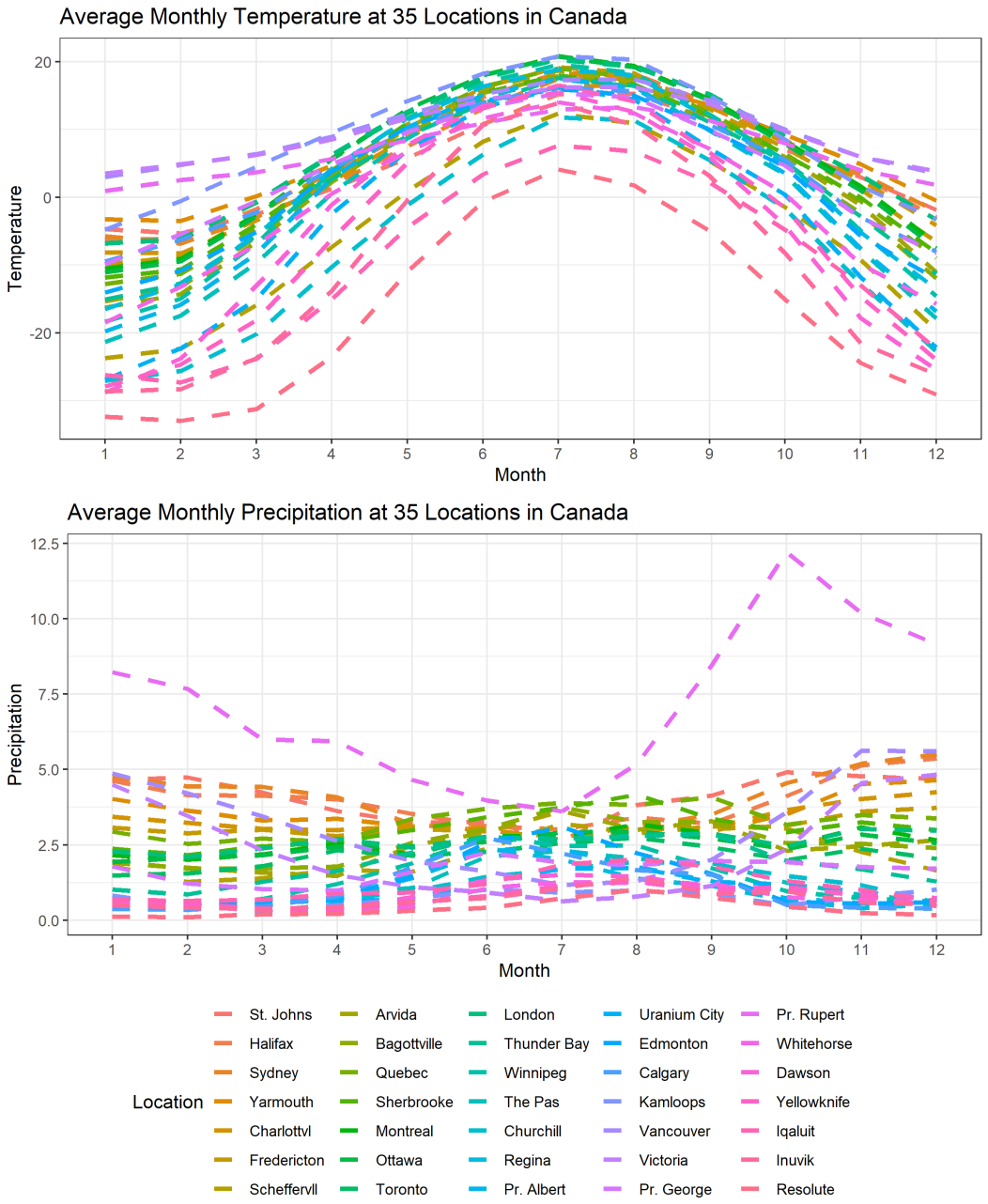


Figure 2.1: Canadian average annual weather cycle data. Average monthly temperature and precipitation at 35 different locations in Canada from 1960 to 1994.

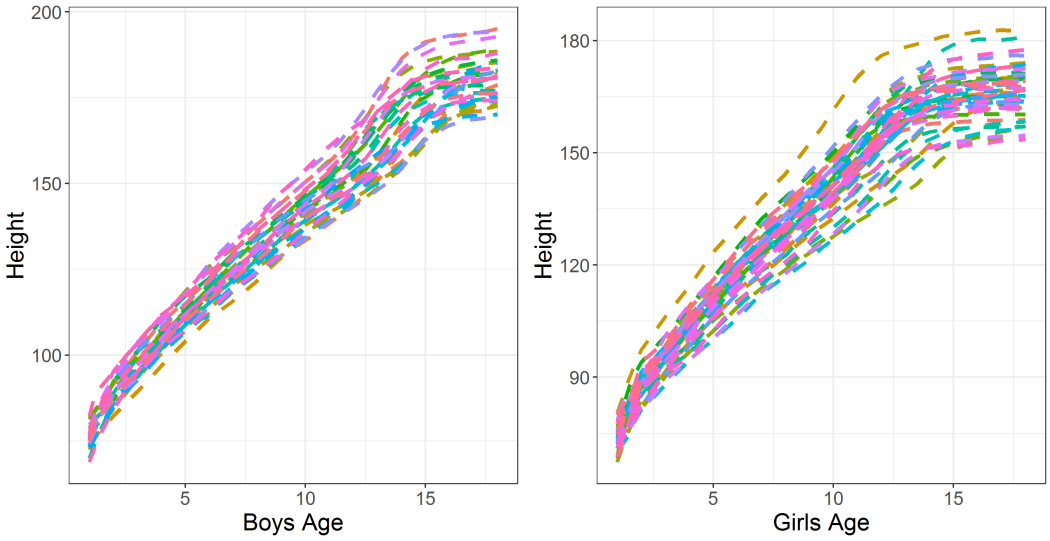


Figure 2.2: Berkeley Growth Study data. Heights of 39 boys and 54 girls from age 1 to 18 and the ages at which they were collected.

From both examples, temperature, precipitation and height govern the behavior of functional variables which are of interest in functional data analysis. By definition, a functional variable is a random process $X(t)$ with t taking values in a closed interval $[t_{min}, t_{max}]$, such that for each fixed t_0, $X(t_0)$ is a random variable. This is the underlying 'smooth' process that generates the data we observe. In other words, since the data are assumed to have been generated from an underlying random process, the set of observations $\{X(t_1), ..., X(t_m)\}$ is considered as a single curve observed at m grid points. This is the main difference between functional and longitudinal data, since in longitudinal analysis we do not assume that such an underlying random process generated the data but instead an $m-$dimensional vector with a specific correlation structure. A simple example of such a process is $X(t) = a_0 + a_1 t$ where a_0 and a_1 are independently and identically distributed $N(0, 1)$ random variables, and $t \in [0, 1]$. For each fixed value of $t = t_0$, $X(t_0)$ is a random

variable with a mean and a variance (both are functions of t_0). Moreover, for any two values t_1 and t_2, $X(t_1)$ and $X(t_2)$ are correlated and their covariance function is defined as $K(t_1, t_2) = \text{Cov}\{X(t_1), X(t_2)\}$.

For each random process, it is possible to have multiple measurements but no parametric assumptions are typically made on the underlying process. The primary interest is to describe the variation of the underlying process. For example, we may ask what feature separates the temperature curves and precipitation curves, how can we discriminate the temperature patterns between Montreal and Resolute, how can we predict a boy's height using girl growth curves, and are growth spurt (rate of change) patterns different for boys and girls.

2.1.2 Basic concepts and notation

In this section, we review some of the essential concepts that define functional data. For simplicity, these concepts will be listed out as follows.

Definition 1 (Derivatives and integrals): Given a function $f(t)$, denote the m^{th} derivative by $f^{(m)}(t) = D^m f = \frac{d^m f(t)}{dt^m}$. Also the integral of f will be denoted by $\int f = \int f(t)dt$.

Definition 2 (Function space): A set of functions which have a particular property in common. For example, the space of all real-valued square integrable functions defined on $[0, 1]$, i.e. $L^2[0, 1]$ space.

Definition 3 (Inner product and norm): For two functions $f(t)$ and $g(t)$ (belonging to the same function space L^2), the inner product is defined as:

$$\langle f, g \rangle = \int f(t)g(t)dt.$$

Given the definition above, the norm of a function is given by:

$$\|f\| = \langle f, f\rangle^{1/2} = \left\{ \int f^2(t)dt \right\}^{1/2},$$

which satisfies the three important properties of a norm:

1. $\|f\| \geq 0$ and $\|f\| = 0$ if and only if $f = 0$
2. $\|af\| = a\|f\|$ for any real number a
3. $\|f + g\| \leq \|f\| + \|g\|$ (Triangle Inequality).

Typically, the norm of a function measures its size and how far from zero the function is in the function space to which it belongs.

Definition 4 (Distance between two functions): The distance between two functions f and g is defined as:

$$d(f, g) = \|f - g\|,$$

which is symmetric and non-negative since it is based on a norm.

Definition 5 (Orthogonality): Two functions f and g are called orthogonal if $\langle f, g\rangle = 0$.

Defintion 6 (Basis expansion of a function): A basis function system for a function space is a set of known (possibly infinitely many) functions $\phi_k, k = 1, 2, ...$ such that any function f can be written as a linear combination of the basis functions, i.e.:

$$f(t) = \sum_{k=1}^{\infty} a_k \phi_k(t),$$

where $a_k, k \geq 1$ are real numbers known as the coefficients of basis representation.

2.1.3 Summary statistics for functional data

Suppose we observe n functional observations $X_1(t), ..., X_n(t)$ observed on $[0, 1]$. The sample mean function is defined as the point-wise average of the observed functions, given by

$$\bar{X}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t),$$

and the sample median function is defined as the point-wise median of the observed functions, given by

$$\bar{X}_m(t) = \text{Med}\{X_i(t), i = 1, ..., n\}.$$

The sample variance function is then naturally derived as

$$\text{Var}_X(t) = \frac{1}{n-1}\sum_{i=1}^{n}\{X_i(t) - \bar{X}(t)\}^2,$$

and the standard deviation function is the square-root of the variance function. Moreover, the covariance function, which characterizes the underlying process that generates the data, is defined as

$$\begin{aligned}\text{Cov}_X(s,t) &= \frac{1}{n-1}\sum_{i=1}^{n}\{X_i(s) - \bar{X}(s)\}\{X_i(t) - \bar{X}(t)\} \\ &= \frac{1}{n-1}\mathbf{X_c}^T\mathbf{X_c},\end{aligned}$$

where $\mathbf{X_c}$ is the centered matrix $\mathbf{X}$ that contains discrete observations of these n functions. Figure 2.3 shows a sample mean and a sample median and function using the temperature data at 35 different locations in Canada. Since the median function is higher than the mean function, this is the temperature

distribution is slightly right-skewed. Figure 2.4 shows a corresponding sample covariance function for these locations. For example, $\text{Cov}_X(1, 2)$ is the covariance of the temperature in 35 locations between January and February. This covariance function has lower covariance toward the center of the heatmap and higher covariance in four corners of the heatmap, indicating that there are more variability of temperature in winter months than summer months.

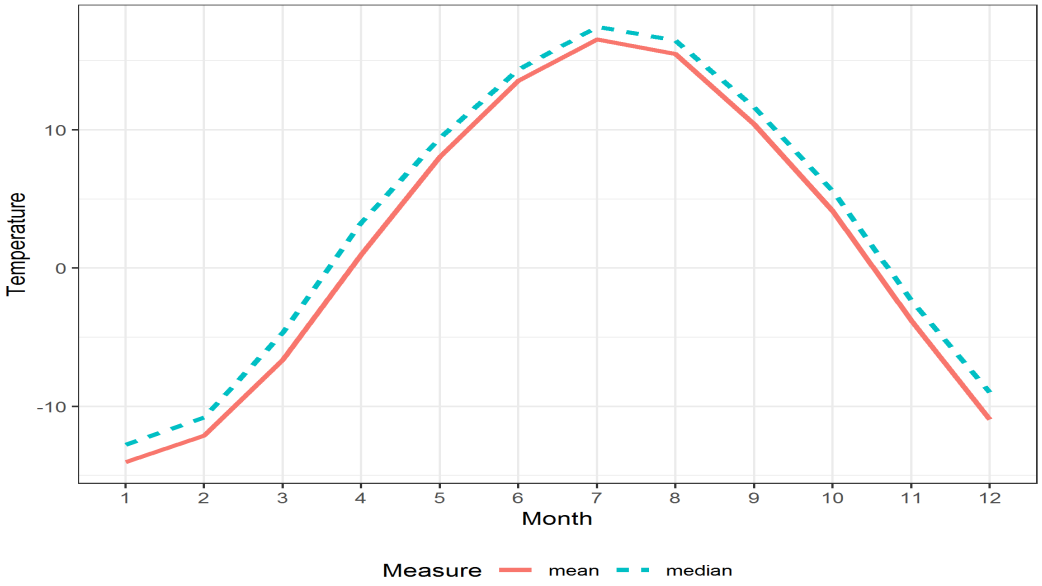


Figure 2.3: Pointwise mean and median functions of the temperature.

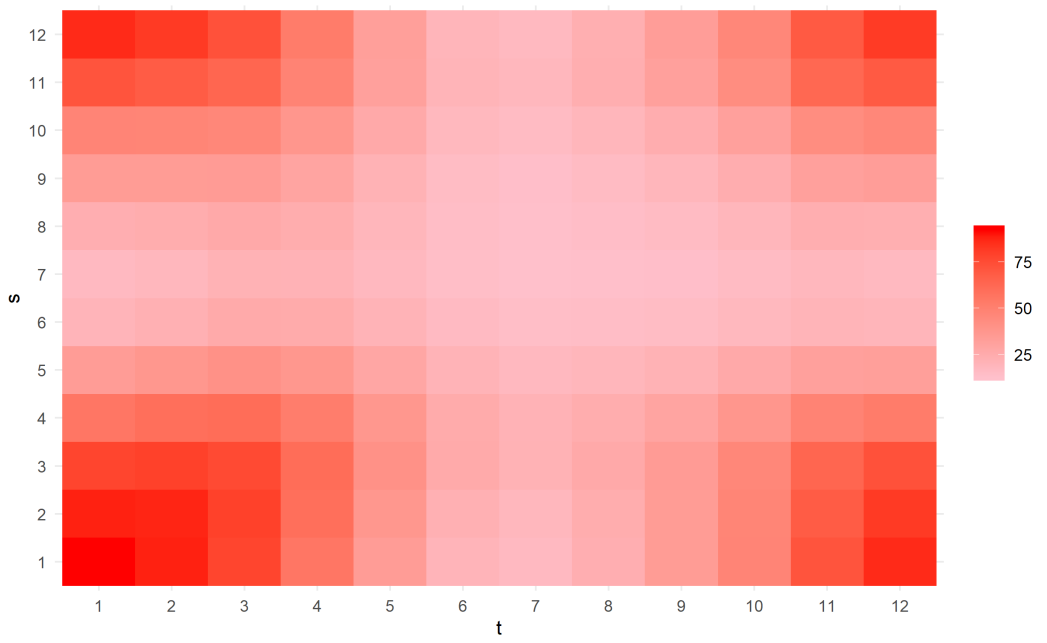


Figure 2.4: Sample covariance heatmap of the temperature.

2.1.4 Challenges of analyzing functional data

When analyzing functional data, we are interested in identifying the features that characterize the functions. Some features may be obvious, such as the sinusoidal shape of the temperature functions in Figure 2.1, but there may be others that are hidden within. Since each function can be considered as an element of an infinite dimensional function space with infinitely many bases, ideally we want to represent each function using only finitely many bases. One approach to accomplish this is the Principal Component Analysis (PCA), which reduces the dimension of the data while explaining a significant percent of variability present in the data. Specifically, given a $p \times p$ covariance matrix,

we seek eigenvectors $\mathbf{u_1}, ..., \mathbf{u_p}$ such that

$$\Sigma = \lambda_1 \mathbf{u_1} \mathbf{u_1^T} + ... + \lambda_p \mathbf{u_p} \mathbf{u_p^T},$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$ are eigenvalues, and the eigenvectors form an orthonormal basis system. Given the data vectors $\mathbf{X_1}, ..., \mathbf{X_n}$, the principal component scores corresponding to the first component are given by

$$PC_{1i} = \mathbf{u_1^T X_i} = \langle \mathbf{u_1}, \mathbf{X_i} \rangle, i = 1, ..., n,$$

and the scores corresponding to the other components are defined similarly. The basic idea of functional PCA is similar. Since the functions are assumed to be generated from an underlying process $X(t)$, we start with the covariance function of this process, $K(s,t) = \text{Cov}\{X(s), X(t)\}$. Thus we seek an eigenfunction decomposition of this covariance function:

$$K(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s) \psi_k(t),$$

where $\phi_k(\cdot), k \geq 1$ are called eigenfunctions, which are known as 'modes of variation', describing a certain percent of variation in the data. These eigenfunctions are denoted as harmonics. Hence, the principal component scores are defined for the first harmonic as:

$$\xi_{1i} = \langle \psi_1, X_i - \bar{X} \rangle = \int \psi_1(t) \{X_i(t) - \bar{X}(t)\} dt,$$

and the scores corresponding to the other components are defined similarly. Later in this chapter, we will discuss in more detail the functional Principal

Component Analysis and apply it extensively in our data analysis.

Predictive models can also be built with functional data. For example, to find out if there is any relationship between the total amount of precipitation of a location and its monthly temperature profile, we can perform functional linear regression with a scalar response and functional covariate. Specifically, let Z_i, $i = 1, ..., 35$ be the total precipitation at the i^{th} location, and let $X_i(t)$, $t = 1, ..., 12$ be the monthly temperature profile. Then the regression model has the form:

$$Z_i = \alpha + \int_1^{12} X_i(t)\beta(t)dt + \epsilon_i,$$

where α is an unknown intercept, $\beta(\cdot)$ is an unknown regression coefficient function and ϵ_i is the error term for each location.

In a different setting, if we want to predict the daily precipitation profile of a location based on its daily temperature profile, we can also fit a functional linear model with a functional response, which has the form:

$$Z_i(t) = \alpha(t) + \int_0^{365} X_i(s)\beta(s,t)ds + \epsilon_i(t),$$

where $Z_i(t)$ is the precipitation at time t, $X_i(t)$ is the temperature at time t, $\alpha(t)$ is an unknown intercept function that describes the overall day-to-day change in precipitation regardless of temperature, $\epsilon_i(t)$ is a random error process, and $\beta(s,t)$ is the unknown regression surface, the relative weight placed on the temperature at day s, for $s = 1, ..., 365$ to predict the precipitation at day t.

2.2 From Discrete to Functional Data

Recall that the philosophy of functional data analysis is to think of observed data functions as single entities, rather than merely as a sequence of individual observations. In practice, functional data are usually observed and recorded discretely as n pairs (t_j, Y_j), where Y_j is a snapshot of the function at time t_j, with possible measurement errors. Therefore, it is crucial to uncover the underlying function for each set of observed discrete data. In this section, we will discuss methods for transforming raw discrete data into smooth functions using linear combinations of basis functions. Indeed, representing data recorded at discrete times as a smooth function would allow us to evaluate the function at any time point, which is extremely useful if we want to compare subjects that were observed at different time points, and examine the rates of change for each underlying curve, given that it is smooth (having one or more derivatives).

In general, observed functional data are recorded as $\{(Y_{ij}; t_{ij}) : j = 1, ..., m_i\}_i$, for $1 \leq i \leq n$, where Y_{ij} is the snapshot of the underlying function at time t_{ij} for subject i, possibly blurred by error, t_{ij} varies in a continuum interval τ and may not be the same across subjects and ϵ_{ij} is the error associated with recording Y_{ij}. The underlying i^{th} function, which is assumed smooth on τ is denoted as X_i and these X_i's are independent realizations of a stochastic process X. In practice, Y_{ij} is assumed to be a measure of X_i at time t_{ij}, i.e. $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$.

2.2.1 Representing Functional Data: Basis Expansions

The goal of basis expansion is to choose a set of bases so that their span includes good approximations of most smooth functions. Since they have to represent the underlying structure in the sample data, they must be able to flexibly exhibit the required curvature where needed, but also to be nearly linear when appropriate. Furthermore, for computational reasons they should be computationally efficient, easy to evaluate, and differentiable as often as required.

A generic underlying function X_i has the form

$$\begin{aligned} X_i(t) &= c_{i1}\phi_1(t) + c_{i2}\phi_2(t) + ... + c_{iK}\phi_K(t) \\ &= \Phi(t)\mathbf{c}_i, \end{aligned}$$

where $\Phi(t) = (\phi_1(t), \ldots, \phi_K(t))$ are predefined basic functions for X_i and $\mathbf{c}_i = (c_{i1}, \ldots, c_{iK})^T$ are coefficients associated with $\Phi(t)$. Here, a basis function system is defined as a set of known functions ϕ_k that are mathematically independent of each other and have the property that any function can be approximated arbitrarily well by taking a weighted sum or linear combination of a sufficiently large number K of these basis functions [Ramsay and Silverman, 2006]. In the example above, we say that $\{\phi_k, k = 1, 2, ..., K\}$ is a basis system for X_i. Since we assume $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$ from above, each 'snapshot' of a function Y_i at time t_{ij} can be estimated as

$$\begin{aligned} Y_i(t_{ij}) &= X_i(t_{ij}) + \epsilon_{ij}, \quad j = 1, \ldots, m_i \\ &\approx c_{i1}\phi_1(t_{ij}) + c_{i2}\phi_2(t_{ij}) + ... + c_{iK}\phi_K(t_{ij}) + \epsilon_{ij}. \end{aligned}$$

Ideally, basis functions should have features that reflect the nature of the data

in order to achieve a good approximation using a small number K of basis functions. Hence, we want to choose an appropriate basis system that only requires a small number of bases to fit the data. Besides reflecting certain characteristics of the data, the small K is more computationally efficient, yields more degrees of freedom for hypothesis testing and confidence intervals, and potentially gives meaningful coefficients that can become interesting descriptors of the data.

For the rest of this section 2.2.1, we will discuss three basis function systems that are widely used in practice and when to use them. To summarize what follows, although a Monomial basis works well with very simple problems, most functional data analyses employ either a Fourier basis for periodic data or a B-spline basis for non-periodic data. Specifically, B-spline bases will be discussed in detail since they are used in the "Application of functional data: the NHANES data analysis" chapter.

Monomial Basis

Polynomials are perhaps the oldest and best known basis function expansion. They can be considered as the senior citizens of the basis world since they can only deal with the simplest functional problems. A polynomial function $X(t)$ has the form

$$X(t) = \sum_{k=1}^{K} c_k t^{k-1},$$

where $t^k, k = 1, \ldots, K$ are the basis functions, i.e. they are the monomials $\{1, t, t^2, t^3, \ldots, t^{K-1}\}$ and c_k are the corresponding coefficients. For simple problems (which usually occur when the function $X(t)$ is smooth), polynomials typically only require $K = 5$ but they have severe problems tracking sharp

localized features, and can run into computational problems for unequally spaced data. Derivative estimation is another limitation of polynomials because their derivatives get simpler as the order of derivative increases, whereas in most real world systems, derivatives become more complex as the order of derivative increases.

Fourier Basis

The Fourier basis system contains basis functions that are sines and cosines of increased frequency:

$$\{1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), ..., \sin(k\omega t), \cos(k\omega t), ..\}.$$

This basis is periodic, and the parameter ω determines the period of oscillation $P = 2\pi/\omega$. The number of bases needed is then $K = 2M + 1$, where M is the largest number of oscillations required in a period of length P.

The Fourier series is a familiar concept to statisticians, engineers and applied mathematicians that possesses a lot of advantages when applied in functional data representation. The Fourier basis functions have excellent computational efficiency, especially if the times of observations are equally spaced due to the orthogonality property of the basis [Ramsay and Silverman, 2006]. It can be shown that if the number of observations is a power of 2 and the arguments are equally spaced, the Fast Fourier transform allows us to find all the coefficients extremely efficiently [Tolimieri et al., 1989]. In terms of fitting data, they are natural for describing periodic data such as the Canadian weather example in Figure 2.1. Their derivatives are also simple to calculate since the derivative of a Fourier series expansion is also a Fourier series expansion, making the Fourier series infinitely differentiable. However, if the data

are known to have discontinuities in the function or in the derivatives, the Fourier series become inappropriate.

B-Spline Basis

Spline functions are the most common choice of approximation system for non-periodic functional data. To define a spline, we first divide the interval over which a function is to be approximated into L sub-intervals, separated by 'knots' (breakpoints). Over each sub-interval, a polynomial of order m, which is the number of constants required in the polynomial (one more than its degree), is fitted and joined with other polynomials from adjacent intervals. Thus, a spline is a piece-wise function made of polynomial segments joined end-to-end such that adjacent polynomials join up smoothly at the knots and are thus differentiable at these points. The number of parameters required to fit a spline function is the order plus the number of interior knots, $m+(L-1)$, which is also the total degrees of freedom. Moreover, derivatives up to order $m-2$ must also match up at these junctions. For example, for the commonly used order four cubic spline, the second derivative is a line and the third derivative is a step function.

Given the definition of a spline function, we can now construct a system of basis spline functions $\phi_k(t)$ which comprise a B-Spline basis. Each basis function $\phi_k(t)$ is a spline function defined by an order m of the polynomial segments and the location of the knots. Specifically, a spline function $S(t)$ with order m and $L-1$ discrete interior knots can be expressed as a linear

combination of $K = m + (L - 1)$ basis functions:

$$S(t) = \sum_{k=1}^{m+L-1} c_k \phi_k(t, \tau),$$

where $\phi_k(t, \tau)$ is the B-Spline basis function defined by the knot sequence τ, evaluated at t and c_k is the associated coefficient. Although there are many ways that such systems can be constructed, [Ramsay and Silverman, 2006] chooses the B-spline basis system developed by [de Boor, 2001], which is the most popular since it allows fast computation for thousands of basis functions and has the flexibility to fit any polynomial of order m. In general, the order of the spline should be at least $m + 2$ if we are interested in m continuous derivatives. The most common choice for polynomial order is $m = 4$ (cubic function), implying continuous second derivatives which are linear functions. Moreover, knots are often equally spaced by default such that each interval contains at least one data point, but it is recommended to place more knots where the function exhibits the most complex variation, and fewer where the function is only mildly nonlinear.

2.2.2 Smoothing Functional Data: Least Squares

Once a basis system is chosen, the natural next step is to quantify the quality of the approximation of the functional data. The classical solution for this problem is least squares estimation, a method that minimizes the sum of squared errors between the observed data and the fitted data. Specifically, let us consider a set of observations $\{(Y_j, t_j) : j = 1, ..., m\}$ and assume $Y_j = X(t_j) + \epsilon_j$, where $X(\cdot)$ is the underlying curve that is observed at the finite grid points t_j

with noise ϵ_j. We want to estimate

$$X(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}^T \Phi(t),$$

where $\mathbf{c}^T = (c_1, ..., c_K)^T$ is the row vector of length K of coefficients and $\Phi(t) = (\phi_1(t), ..., \phi_K(t))$ is the column vector of length K containing basis functions. Let Φ be the $m \times K$ matrix obtained by row-stacking $\Phi(t_j)$ for $j = 1, ..., m$ and $\mathbf{Y} = (Y_1, ..., Y_m)^T$. The ordinary least square (OLS) criterion assumes that residuals are independently and identically normal with mean 0 and variance σ^2, i.e: $\epsilon_j \overset{iid}{\sim} N(0, \sigma^2)$. Given that this assumption is appropriate for the data, we can determine the coefficients of the expansion $\mathbf{c}^T = (c_1, ..., c_K)^T$ by minimizing

$$\begin{aligned} SSE(\mathbf{c}) &= \sum_{j=1}^{m} \{Y_j - X(t_j)\}^2 \\ &= \sum_{j=1}^{m} \{Y_j - \mathbf{c}^T \Phi(t_j)\}^2 \\ &= (\mathbf{Y} - \Phi\mathbf{c})^T (\mathbf{Y} - \Phi\mathbf{c}). \end{aligned}$$

After taking the derivative of SSE($\mathbf{c}$) with respect to $\mathbf{c}$ and solving, the OLS estimate of $\mathbf{c}$ is therefore

$$\hat{\mathbf{c}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{Y},$$

and the fitted value at time t_j is

$$\begin{aligned} \hat{X}(t_j) &= \Phi(t_j)\hat{\mathbf{c}} \\ &= \Phi(t_j)(\Phi^T \Phi)^{-1} \Phi^T \mathbf{Y}, \quad j = 1, ..., m. \end{aligned}$$

In practice, the independently and identically distributed residual assumption of OLS may not be met. For example, the data may be uncorrelated but exhibit heteroskedastiscity (non constant variance). In such situations, we weight the observations by extending the least squares criterion to the weighted least squares (WLS) form:

$$\begin{aligned} WMSE(\mathbf{c}) &= \sum_{j=1}^{m} w_j \{Y_j - X(t_j)\}^2 \\ &= \sum_{j=1}^{m} w_j \{Y_j - \mathbf{c}^T \Phi(t_j)\}^2 \\ &= (\mathbf{Y} - \Phi\mathbf{c})^T \mathbf{W} (\mathbf{Y} - \Phi\mathbf{c}), \end{aligned}$$

where $\mathbf{W}$ is a diagonal matrix with the diagonal elements equal to w_j. It is usually estimated by the covariance matrix $\mathbf{\Sigma}_\epsilon$ of $\mathbf{Y}$ as $\mathbf{W} = \mathbf{\Sigma}_\epsilon^{-1}$, where since the Y_j are uncorrelated ($\text{cor}(Y_i, Y_j) = 0$ for $i \neq j$), the off-diagonal terms of $\mathbf{\Sigma}_\epsilon$ are zeroes. The weighted least squares estimate $\hat{\mathbf{c}}$ of the coefficient vector $\mathbf{c}$ is then

$$\hat{\mathbf{c}} = (\Phi^T \mathbf{W} \Phi)^{-1} \Phi \mathbf{W} \mathbf{Y}$$

and the fitted curve at point t_j is calculated as

$$\hat{X}(t_j) = \Phi(t_j)(\Phi^T \mathbf{W} \Phi)^{-1} \Phi \mathbf{W} \mathbf{Y}, \quad j = 1, ..., m.$$

2.2.3 Choosing the number of basis functions

During the data fitting process, the more basis functions we select, the better the fit to the data but the higher the risk of fitting noise or variation that we do not need. Although the bias would be small, the sampling variance would be large, similar to the over-fitting problem in linear regression. Nevertheless, if

we do not choose enough basis functions, we may miss some important aspects of the function that we are trying to capture, potentially resulting in large bias. Hence, the mean squared error is often used as a loss function that controls the bias and variance of the estimator of the curve. Specifically, for a fixed t, the mean squared error (MSE) in estimating $X(t)$ is defined as

$$\begin{aligned} MSE\{\hat{X}(t)\} &= E[\{\hat{X}(t) - X(t)\}^2] \\ &= \text{Bias}^2\{\hat{X}(t)\} + \text{Var}\{\hat{X}(t)\}, \end{aligned}$$

where the bias of the estimator is

$$\text{Bias}\{\hat{X}(t)\} = X(t) - E\{\hat{X}(t)\}$$

and the corresponding sampling variance is

$$\text{Var}\{\hat{X}(t)\} = E[\{\hat{X}(t) - E[\hat{X}(t)]\}^2].$$

The optimal number of basis functions K to fit a curve would minimize the integrated mean squared error

$$MSE(\hat{X}) = \int_{\tau} MSE\{\hat{X}(t)\}dt,$$

which can be be approximated numerically by $\frac{1}{m}\sum_{j=1}^{m} MSE\{\hat{X}(t_j)\}$ when the number of time points m is large.

2.3 Principal components analysis for functional data

After the preliminary steps of registering and displaying the data, we want to explore the data to see the features characterizing typical functions. Some of these features can be detected easily, such as the sinusoidal nature of the temperature curves, but other features might be more obscure. Principal components analysis (PCA) of functional data is then a key technique to identify these hidden features. In fact, it is the first method to be considered in the early literature of functional data analysis since it provides us a way to examine the variance-covariance and correlation functions that can be very informative. In this section, we will briefly review the classical principal components analysis for multivariate data and then introduce its version for functional data.

2.3.1 PCA for multivariate data

One of the problems with multivariate data is that there are simply too many variables to make the application of graphical techniques successful in providing an informative initial assessment of the data. Moreover, having too many variables can also cause problems, such as multicollinearity, for other multivariate techniques that the researcher may want to apply to the data. Principal components analysis is a multivariate technique with the central aim of reducing the dimensionality of a multivariate data set while accounting for as much of the original variation as possible. This aim is achieved by creating a new set of variables, the principal components, that are linear combinations of the original variables, which are uncorrelated and are ordered so that the first few

of them account for most of the variation in all the original variables. Ideally, the result of a principal components analysis would be the creation of a small number of new variables that can be used as surrogates for the originally large number of variables and consequently provide a simpler basis for graphing or summarising the data, and for further multivariate analyses of the data.

Let the data matrix $\mathbf{X}$ be of size $n \times p$, where n is the number of samples and p is the number of variables. Let us also assume that each variable is centered, i.e. column means have been subtracted and the centered means are now equal to zero. Principal components analysis describes variation in a set of correlated variables, $\mathbf{x}^T = (x_1, ..., x_p)^T$, in terms of a new set of uncorrelated variables, $\mathbf{y}^T = (y_1, ..., y_p)^T$, each of which is a linear combination of the $\mathbf{x}$ variables. The new variables are derived in decreasing order of 'importance' in the sense that y_1 accounts for as much of the variation as possible in the original data amongst all linear combinations of $\mathbf{x}$. Then y_2 is chosen to account for as much of the remaining variation as possible, subject to being uncorrelated with y_1, and so on. The new variables defined by this process, $y_1, ..., y_p$, are the orthogonal principal components. The general hope of principal components analysis is that the first few components will account for a substantial proportion of the variation in the original variables, $x_1, ..., x_p$, and can be used to provide a convenient lower-dimensional summary of these variables.

Finding the principal components

The first principal component of the observations, y_1, is the linear combination

$$y_1 = \psi_{11}x_1 + \psi_{12}x_2 + ... + \psi_{1p}x_p,$$

whose sample variance is greatest among all such linear combinations. Since the variance of y_1 could be increased without limit simply by increasing the coefficients $\boldsymbol{\psi}_1^T = (\psi_{11}, \psi_{12}, ..., \psi_{1p})^T$, a normalization restriction must be placed on these coefficients: the sum of squares of the coefficients should take the value 1. Hence, to find the coefficients defining the first principal component, we need to choose the elements of the vector ψ_1 that maximize the variance of y_1, subject to the constraint $\boldsymbol{\psi}_1^T \boldsymbol{\psi}_1 = 1$. The second principal component, y_2, is defined to be the linear combination

$$y_2 = \psi_{21} x_1 + \psi_{22} x_2 + ... + \psi_{2p} x_p$$

that has the greatest variance subject to the following two conditions: $\boldsymbol{\psi}_2^T \boldsymbol{\psi}_2 = 1$ and $\boldsymbol{\psi}_2^T \boldsymbol{\psi}_1 = 0$, i.e. y_1 and y_2 are uncorrelated. Continuing in this fashion, the j^{th} principal component is that linear combination $y_j = \boldsymbol{\psi}_j^T \mathbf{x}$ that has the greatest variance subject to the conditions $\boldsymbol{\psi}_j^T \boldsymbol{\psi}_j = 1$ and $\boldsymbol{\psi}_j^T \boldsymbol{\psi}_i = 0$ for all $i < j$. To calculate each $\boldsymbol{\psi}_j$, we solve the eigenequation

$$\boldsymbol{\Sigma} \boldsymbol{\psi_j} = \lambda_j \boldsymbol{\psi_j},$$

where $\boldsymbol{\Sigma}$ is the sample variance-covariance matrix which is defined as $\boldsymbol{\Sigma} = N^{-1} \mathbf{X}^T \mathbf{X}$ ($\mathbf{X}$ is the centered data matrix), $\boldsymbol{\psi_j}$ is the j^{th} eigenvector of $\boldsymbol{\Sigma}$ with the corresponding eigenvalue λ_j. Putting all eigenvectors as columns of a matrix $\mathbf{V}$ and corresponding eigenvalues as entries of a diagonal matrix $\boldsymbol{\Lambda}$, the above equation can be extended to $\boldsymbol{\Sigma}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}$, or $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, the eigen decomposition of $\boldsymbol{\Sigma}$. Here, the columns of $\mathbf{V}$ are called the principal components (PCs) which are orthogonal with unit norm; $\boldsymbol{\Lambda}$ is a diagonal matrix, defined

as $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$ where the entries are non-negative and arranged in decreasing order. The entry λ_k, $k = 1, \ldots, p$ gives the variance of the data along the corresponding PC and the proportion of variance explained by the k^{th} PC is defined as $\lambda_k / \sum_{l=1}^{p} \lambda_l$. Finally, the projections of the data on the principal components are known as PC scores; these can be seen as new transformed variables. The j^{th} principal component projection is given by the j^{th} column of XV and the coordinates of the i^{th} data point in the new PC space are given by the i^{th} row of XV.

2.3.2 PCA for functional data

Functional principal components analysis (fPCA) was first developed by C. Radhakrishna Rao in 1958 [Rao, 1958]. It is used to analyze the geometry of the functions, capture the principal modes of variation and reduce the dimension of the data. Let $X_1(t), \ldots, X_n(t)$ denote independent and identically distributed random functions on a compact interval $\mathcal{T}$ such that each function $X_i(t)$ belongs to the functional space of all real valued square integrable functions defined on $[\mathcal{T}]$, i.e. the $L^2[\mathcal{T}]$ space, with the true mean function defined as $\mu(t) = E[X_i(t)]$ and their corresponding covariance function defined as $\Sigma(s,t) = \text{Cov}\{X_i(s), X_i(t)\}$. For simplicity, let us assume that the functions are observed fully on $\mathcal{T}$ and without noise.

Similar to PCA, fPCA is based on finding principal component scores of maximum variance that highlight features of the smooth underlying curves. Specifically, to find the first functional principal component, we find the principal component weight function (eigenfunction) $\psi_1(t)$ for which the set of

values

$$\begin{aligned}\xi_{1i} &= \int_{\mathcal{T}} \psi_1(t) X_i(t) dt \\ &= \langle \psi_1, X_i \rangle \quad i = 1, \ldots, n\end{aligned}$$

has the largest variance, subject to the constraint $\langle \psi_1, \psi_1 \rangle = 1$. The second functional principal component finds the eigenfunction $\psi_2(t)$ for which the set of values

$$\begin{aligned}\xi_{2i} &= \int_{\mathcal{T}} \psi_2(t) X_i(t) dt \\ &= \langle \psi_2, X_i \rangle \quad i = 1, \ldots, n\end{aligned}$$

has the largest variance, subject to the constraint $\langle \psi_2, \psi_2 \rangle = 1$ and $\langle \psi_1, \psi_2 \rangle = 0$. Continuing in this fashion, the k^{th} functional principal component score finds the eigenfunction $\psi_k(t)$ for which the set of values

$$\begin{aligned}\xi_{ki} &= \int_{\mathcal{T}} \psi_k(t) X_i(t) dt \\ &= \langle \psi_k, X_i \rangle \quad i = 1, \ldots, n\end{aligned} \tag{2.1}$$

has the largest variance, subject to the constraint $\langle \psi_k, \psi_k \rangle = 1$ and $\langle \psi_k, \psi_j \rangle = 0$ for all $j < k$.

In order to calculate the eigenfunctions $\{\psi_k(t), k = 1, \ldots, n\}$ and represent any given function $X(t)$ in terms of these eigenfunctions, we use the theories from Mercer's theorem and the Karhunen-Loève expansion [Happ and Greven, 2015]. Mercer's theorem allows for the eigen-decomposition of a covariance function $\Sigma(s,t)$ into eigenvalues λ_k and eigenfunctions $\psi_k(t)$. Under the assumption that $\Sigma(s,t)$ is defined continuously over the compact

interval $\mathcal{T}$ and square integrable, Mercer's theorem states that there exists an orthonomal sequence ψ_k of continuous functions in $L^2[\mathcal{T}]$ with unit norm and a non-increasing sequence of positive numbers $\lambda_1 \geq \lambda_2 \geq ... > 0$ such that

$$\Sigma(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t) \quad s,t \in \mathcal{T},$$

with the eigenvalues and eigenfunctions being solutions to

$$\int_{\mathcal{T}} \Sigma(s,t)\psi_k(s)ds = \lambda_k \psi_k(t).$$

A complete proof of this theorem can be found in [Bosq, 2000]. When Mercer's theorem holds, the Karhunen-Loève theorem states that using the basis functions determined by the eigenfunctions of the covariance function, the curves X_i have the following representation

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t),$$

where the basis coefficients are the principal component scores ξ_{ik} defined similarly as in equation 2.1:

$$\xi_{ik} = \int_{\mathcal{T}} \{\psi_k(t)(X_i(t) - \mu(t))\}dt \tag{2.2}$$

such that $\xi_{ik} \sim N(0, \lambda_k)$ and they are uncorrelated for different k. Recall that we are interested in finding the set of K orthogonal functions $\{\psi_1, \ldots, \psi_k\}$ for which if $\hat{X}_i(t)$ denotes the best approximation of $X_i(t)$ using these basis

functions, then the mean integrated squared error (MISE) criterion

$$\begin{aligned} MISE &= \sum_{i=1}^{n} ||X_i - \hat{X}_i||^2 \\ &= \sum_{i=1}^{n} \int_{\mathcal{T}} \{X_i(t) - \hat{X}_i(t)\}^2 dt \end{aligned} \tag{2.3}$$

is minimized. [Ramsay and Silverman, 2006] show that the set of basis functions that minimizes equation 2.3 has the additional property that it maximizes the amount of variation explained in the random functions $X_i(t)$. Hence, the collection of the first K eigenfunctions in the sample of curves $\{X_i(t), i = 1, \ldots, n\}$ forms a set of basis functions that minimizes the above MISE criterion. Since these basis functions are derived directly from the functional data instead of being chosen like the Fourier or B-spline basis, they can be considered as empirical basis functions.

Chapter 3

Application of functional data: the NHANES data analysis

3.1 Introduction

The National Health and Nutrition Examination Survey (NHANES) is a cross-sectional, nationally representative survey designed to evaluate the health and nutritional status of adults and children in the United States [CDC, 2016]. The survey samples around 5000 non-institutionalized civilians annually to represent the US population. In particular, NHANES oversamples underrepresented groups, including elderly people 60+ years old, African Americans, Asians, and Hispanics. The survey involves a 4-stage process to sample participants, which indicates that the sample is not a simple random sample from the US population. To make the sample representative for the US population each individual sampled in the NHANES has a survey weight, which is defined as the number of individuals in the US population represented by that individual. These survey weights need to be incorporated in any analysis to ensure

that results are generalizable to the US population. The survey collects demographic, socioeconomic, dietary, and health-related information through home interviews, and medical, dental, and physiological measurements through physical examinations in mobile centers [CDC, 2016]. Moreover, NHANES started to monitor participants' physical activity using an accelerometer during a 1-week study for its 2003-2004 and 2005-2006 cohorts. The National Center for Health Statistics also provides a mechanism for linking NHANES cohorts with death certificate records from the National Death Index (NDI) [NCHS, 2015]. This allows us to investigate the associations between participants' activity and other non-activity related characteristics and future mortality.

For our research, we are interested in: 1) exploring the associations between participants' physical activity, demographic, and health-related characteristics and 5-year all-cause mortality; 2) identifying the ranking of the most predictive predictors and their relative effects on mortality; 3) comparing derived measures of physical activity (PA) to established predictors of 5-year all-cause mortality.

3.2 Study Population

The NHANES is a large study conducted by the Centers for Disease Control (CDC) to assess the health and nutritional status of the US population [CDC, 2016]. These data include: (1) responses to demographic, socioeconomic, and health related survey questions; (2) medical, dental, physiological examination, and clinical laboratory tests; and (3) PA information measured by accelerometers. Non-institutionalized civilian residents of the United States were selected to participate in this study according to the CDC sam-

ple design specifications [Curtin et al., 2013]. Each study participant was assigned a survey weight equal to the number of people he or she represents in the US population. The NHANES 2003-2004 and 2005-2006 data were downloaded, processed, and combined with survey weights and mortality data (updated through 2015). Data are organized in the **R** package **rnhanesdata** [Leroux et al., 2019].

The NHANES 2003-2004 and 2005-2006 have a total of 14,631 participants with accelerometry data. For this analysis, we excluded participants who: (1) were younger than 50 years of age, or 85 and older at the time they wore the accelerometer (10,859 participants); (2) had missing BMI or education predictor variables (41 participants); (3) had fewer than 3 days of data with at least 10 hours of estimated wear time or were deemed by NHANES to have poor quality data (517 participants); non-wear periods were identified as intervals with at least 60 consecutive minutes of zero activity counts and at most 2 minutes with counts between 0 and 100; (4) had missing mortality information (21 participants); (5) had missing systolic blood pressure (SBP), total or HDL cholesterol measurements (293 participants). Among the remaining participants, 86 did not have alcohol consumption information and were retained in the dataset by introducing the category 'Missing Alcohol'. The final dataset contained 2,978 participants with 297 deaths in the first five years after the accelerometer study.

3.3 Variables

3.3.1 Traditional mortality predictors

We integrated the NHANES data with the US national mortality registries starting with the socio-demographic factors age, sex, race/ethnicity, and educational attainment. In NHANES, race/ethnicity was coded as Non-Hispanic White (White), Mexican American (Mexican), Non-Hispanic Black (Black), Other Hispanic and Other. Educational attainment was coded as less than high school, high school equivalent and greater than high school. We further included smoking status (never, former, current), alcohol consumption (non-drinker, moderate drinker, heavy drinker, missing alcohol), body mass index (BMI; kg/m2), mobility difficulty (yes/no), diabetes, coronary heart disease (CHD), congestive heart failure (CHF), stroke, cancer, systolic blood pressure (SBP), total cholesterol (mg/dL), and HDL cholesterol (mg/dL). Mobility difficulty was defined as a positive response to any of the following questions: (1) difficulty walking a quarter mile; (2) difficulty climbing 10 stairs; or (3) use of any special equipment to walk.

3.3.2 Accelerometry derived predictors

According to the NHANES protocol, the minute-by-minute activity data was recorded using a hip-worn ActiGraph AM-7164 (formerly the CSA/MTI AM-7164) accelerometer, as shown in Figure 1.3. Each participant was instructed to wear the device for a period of 7 consecutive days from the day of NHANES examination and remove the device during sleep and water-related activity, such as swimming and bathing. The device was returned to the CDC by mail

in postage-paid padded envelopes. Not every study participant wore the device for the full 7-day period.

The high volume of minute-level activity measurements is challenging, which is why the current practice is to take summary measures. Popular PA summaries based on actigraphy include: (1) total activity count (TAC); (2) total log(1+activity count), referred to as total log activity count (TLAC); and (3) total minutes of moderate/vigorous physical activity (MVPA), where MVPA is defined as the total time with more than 2020 counts per minute. While informative, these summaries do not reflect the full complexity of daily activity patterns and may miss important information that could be associated with health and functional status. To evaluate the effect of daily PA patterns on mortality we introduce 12 additional summary variables (TLAC 12AM-2AM, TLAC 2AM-4AM,..., TLAC 10PM-12AM), where each variable corresponds to the total log(1+activity count) in a 2-hour interval. For example, TLAC 12AM-2AM is the total log activity between 12AM and 2AM. We also used two measures of activity fragmentation: transition probabilities from sedentary to active (SATP) and active to sedentary (ASTP) [Di et al., 2017]. The sedentary to active transition probability (SATP) is defined as SATP=n_s/T_s, where n_s is the total number of sedentary bouts (periods where the activity count is less than 100) and T_s is the total sedentary time. Specifically, if the duration of the longest sedentary bout is denoted by D_s, the number of bouts of length t is denoted by $n_s(t)$, then the total sedentary time can be represented as $T_s = \sum_{t=1}^{D_s} n_s(t) \times t$, and the total number of sedentary bouts can be represented as $n_s = \sum_{t=1}^{D_s} n_s(t)$. SATP is inversely proportional to the average length of the inactivity bout, but has better statistical properties (e.g., symmetric distribution with normal tails across

study participants). Similarly, the active to sedentary transition probability (ASTP) is defined as ASTP=n_a/T_a, where n_a is the total number of active bouts and T_a is the total activity time. Larger values of transition probabilities correspond to shorter average bout duration and more frequent switching between states and more fragmented PA. The total accelerometer wear time (Wear Time in minutes) and total sedentary/sleep/non-wear time were also included. Since each participant had 3-7 days of available accelerometry data, for each accelerometry derived summary measure we calculated the measure for each day (i.e., TAC_day1, TAC_day2,..., TAC_day7) and then averaged across available days.

We also propose to use principal component analysis (PCA) to derive additional predictors. This is a widely accepted and fast approach to addressing whether something has been missed by simple summaries of the data. PCs were obtained as follows: (1) transform minute-level activity count data as x $\rightarrow$ log (1 + x), where x is the minute level activity count; (2) arrange all activity trajectories into a 18373 by 1440 dimensional matrix, X, where each row corresponds to a subject/day and each column corresponds to a specific minute (time) of the day; (3) conduct functional PCA (fPCA) on the matrix X using the **fpca.face()** function [Xiao et al., 2014] in the **refund** package [Goldsmith et al., 2018] in **R**; (4) retain the first 6 PCs (shown in Figure 3.1), which explain 57% of the variability in the activity data; and (5) obtain the score for each day on each PC and calculate the mean and standard deviation of these scores for each subject across days. More specifically, let z_{ijk} be the score for subject i, on day j and PC k. Then we construct the following

additional $2K$ variables (2 for each of the first $K = 6$ PCs):

$$m_{ik} = \bar{z}_{ik} = \frac{1}{J_i}\sum_{j=1}^{J_i} z_{ikj}, \quad i = 1, ..., N \quad j = 1, ..., J_i \quad k = 1, ..., 6$$

$$s_{ik} = sd(z_{ik}) = \sqrt{\frac{\sum_{j=1}^{J_i}(z_{ijk} - \bar{z}_{ik})^2}{J_i - 1}}, \quad i = 1, ..., N \quad j = 1, ..., J_i \quad k = 1, ..., 6.$$

To see what PCA-derived variables to retain we started with a model containing the standard demographic, behavioral and comorbidity variables and conducted forward selection on the means and standard deviations of the scores on the 6 PCs (a total of 12 variables). Using this procedure, the average scores for the first (m_{i1}; odd ratio (OR) = 1.014, CI: (1.004, 1.026); p = 0.008), and the standard deviation of the sixth PCs (s_{i6}; OR = 0.926, CI: (0.888, 0.965); p <0.001) were found to be statistically associated with 5-year all-cause mortality.

PCA is widely used, but is often criticized for the lack of intuition and transportability potential across studies. To overcome this problem, we have inspected the PCs and replaced them with surrogate variables that are intuitive and can be calculated directly from the observed data.

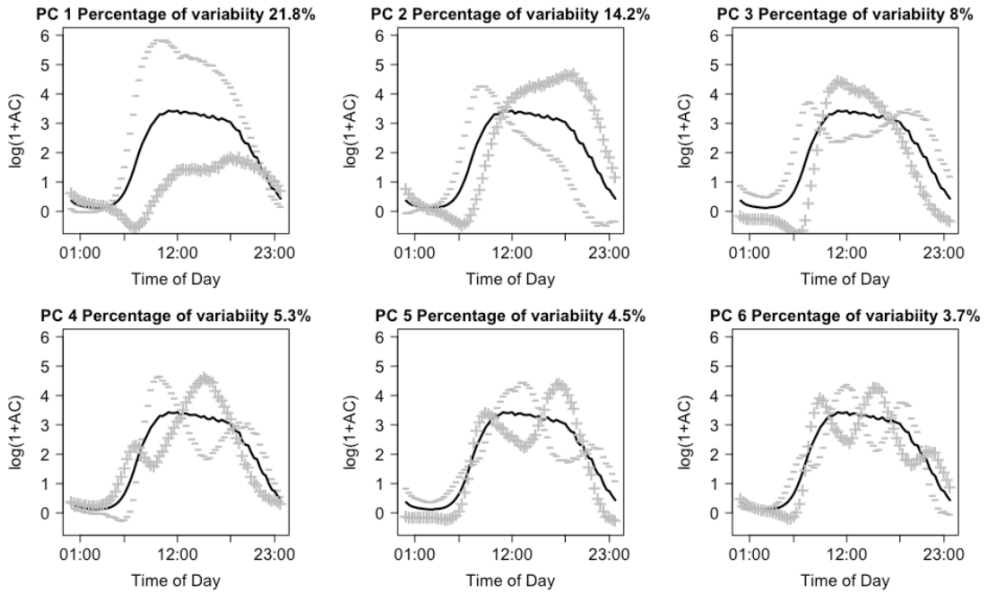


Figure 3.1: First 6 principal components calculated on the population, minute level NHANES accelerometry data. Solid lines represent the population average curve; +,- lines denote the effect of being 2 standard deviations from a score of 0 on the particular principal component.

Identifying potential surrogate measures is based on (subjective) interpretations involving the shapes of each principal component. Fundamentally, the idea is to use visual inspection of the principal components to identify the 'dominant' features of each component. Then, we return to the original data, and calculate a statistic which we believe captures this dominant feature. For example, looking at the upper-left panel of Figure 3.1, we see that days which load negatively on the first principal component tend to be extremely active, while those who load positively tend to be very inactive. As a result, one reasonable 'guess' at a surrogate measure which is highly associated with average first component is simply the total log transformed activity count (TLAC) for that day. If that is true, we would also expect that the average PC1 score

within subjects is highly correlated with their average TLAC across days. In our data, average TLAC and average PC1 score are highly (negatively) correlated ($\hat{\rho} = -0.87$), which is expected based on the sign of the first PC.)

This procedure would then be repeated for each feature identified as potentially predictive. In our application, we are also interested in the standard deviation of PC6. Looking at the bottom-right panel of Figure 3.1, we see that there are 6 periods where the contrast is highest between days with positive and negative loadings (i.e. the difference between the + and - curves is largest). One reasonable guess for a statistic which is highly correlated with PC6 score is the difference in average activity during the specific time periods where positive/negative loadings are high/low, respectively. For example, days that load highly on PC6 should, on average, have higher activity during the mid morning (8AM-10AM), late afternoon (3PM-5PM) and late evening (10PM-12AM) and lower activity during the early morning (5AM-7AM), late morning/early afternoon (11AM-1PM), and early evening (6PM-8PM).Since we are interested in the standard deviation of PC6 score, we calculate the standard deviation of average log-transformed activity counts during these periods as a surrogate measure for s_{i6}. In our analysis, we used all 6 of these time periods and obtained an observed correlation of $\hat{\rho} = 0.87$, though multiple choices can be explored to see which statistic has the highest correlation with s_{i6}.

3.3.3 Intuition behind fPCA

A major problem with PC analysis is that it is not always intuitive and requires a degree of familiarity with matrix algebra and complex trajectories

(functional data analysis in statistics speak). While some of these problems are unavoidable given the complexity of the data, we will now provide the needed intuition for understanding both the PCs and the implication of our findings on the original data scale (daily minute-level activity profiles). We will start by explaining the first 2 PCs, which are shown in the left panel of Figure 3.2.

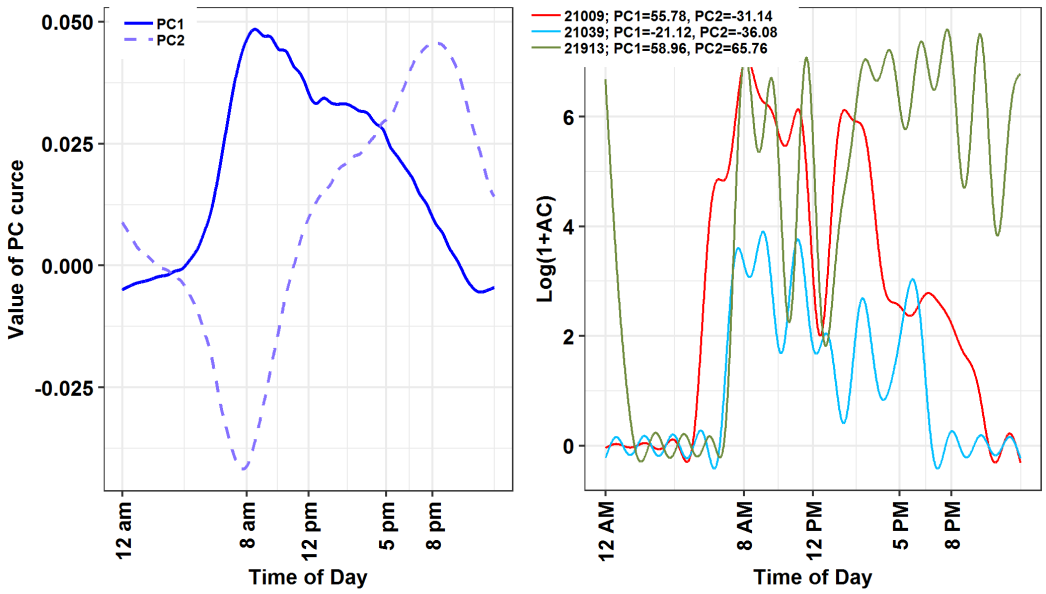


Figure 3.2: Left panel: the first two principal components that explains the overall variability in the observed daily profiles of activity. The x-axis shows the time of day and the y-axis shows the values of PC curve. Individuals with a positive score on the first PC on a given day will tend to have less activity during the night hours and more activity during the day hours than the average activity across all subject-days. The second PC reflects the contrast between morning and afternoon activity. Right panel: Examples of activity profile for 3 subjects to show the connection between the PCs and the activity profiles. The x-axis shows the time of day and the y-axis shows log(1+AC) values.

The first PC (solid blue line) captures 21.8% of the overall variability in the observed daily profiles of activity. It has a distinct shape, with values starting negative between 12AM and 5AM then becoming strongly positive with a peak around 8AM and slowly decreasing but staying positive until 9:30PM, and then becoming negative after 9:30PM. This is exactly what we expected to see. For each subject, we have 3 - 7 days of valid activity data. Individuals with a positive score on this component (a.k.a., positively loaded on the first PC) on a given day will tend to have less activity during the night hours and more activity during the day hours than the average activity across all subject-days. The biggest difference between such a subject's day and the average daily activity across all subjects is centered on the morning hours (8AM-9AM). In contrast, the second PC (dashed line) captures 14.2% of the overall variability in the observed daily activity profiles. It starts positive between 12AM and 2AM, then becomes negative between 2AM and 11AM, with a negative peak at 8AM, increases between 11 AM and 8PM, and decreases while staying positive after 8PM. Participants with positive scores on this component will be more active in the evening and less active in the morning than the average individuals' daily activity.

Now, let us investigate in detail the connection between PCs, scores and individual daily trajectories. The right panel in Figure 3.2 displays one day of activity profiles for 3 subjects. Here, we plot the activity data smoothed for each subject and day using a thin-plate penalized regression spline with 30 knots as implemented using the **gam()** function in the **mgcv** package in **R**. The individual days for subjects 21009 (red line) and 21913 (green line) have overall high activity, which is reflected by the high positive loadings on PC 1 (56.27 and 58.67, respectively). In contrast, the individual day's activity for

subject 21039 (blue line) has lower levels, which is reflected by the negative score on the first PC (-20.82). Subjects 21009 (red line) and 21039 (blue line) are mostly active between 7AM and 9PM and their corresponding scores of the second principal component are negative (-30.83 and -36.36, respectively). Subject 21913 (green line) is however, unusually highly active during night hours (8PM - 2AM) with a highly positive score on the second PC (66.54).

The last, but not least important interpretation is of means of scores versus standard deviation of scores. In Figure 3.2, we have displayed three days, one for each subject. However, each subject has multiple days and each day will get a score and a pattern. For example, on days when subject 21009 (red line) is less active the score on PC1 will be lower, even if the general pattern stays the same. Thus, for every subject and PC we obtain a vector of scores; for 21009 (red line) we obtain (38.39, 56.27, 35.55, 60.61, 40.91, 48.67, -21.25) on the first PC, where we showed only the trajectory corresponding to the 2^{nd} day. What we calculate is the mean of these scores, 37.02, and the standard deviation, 27.32. The mean of the scores is relatively easy to interpret, as it represents whether the average of the 7 days is higher or lower on PC1. The standard deviation captures the day-to-day variability of the individual. In this case the mean and standard deviation for subject 21009 (red line) were 37.02 and 27.32, respectively. In contrast, for subject 21039 (blue line), the mean and standard deviation were -19.16 and 5.02, respectively, both much smaller than for subject 21009 (red line). This means that both the overall mean and the daily variability around this larger mean are larger for subject 21009 than for subject 21039 (blue line). This is depicted in Figure 3.3, where we show smoothed activity data for all days for subjects 21009 (left panel) and 21039 (right panel). Note that the red lines tend to be higher the blue lines and

that the blue lines are less variable around their means. We conclude that, in general, average PC1 scores will tend to distinguish between lower and higher activity individuals whose are, on average, more active over the course of days with available activity data. In contrast, average PC2 scores will distinguish between individuals who, on average, have high activity in the morning and low in the evening/night and individuals who, on average, have lower activity intensity in the morning and higher during the evening/nighttime.

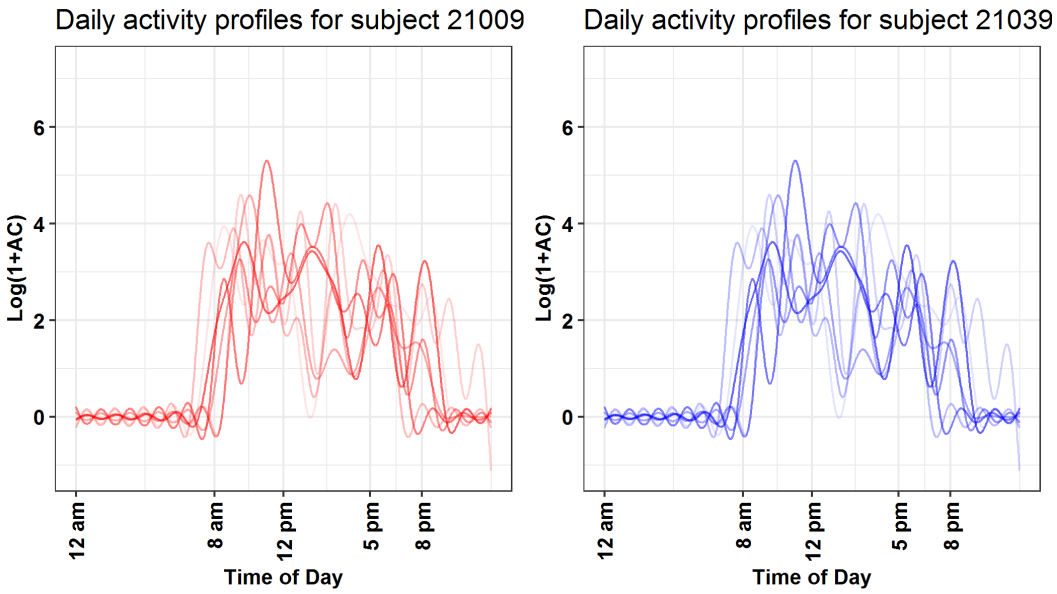


Figure 3.3: Left panel: Daily activity profile for subject 21009. Right panel: Daily activity profile for subject 21039. For both panels, the x-axis shows the time of day and the y-axis shows log(1+AC) values. This figure demonstrates the day-to-day variability of the activity profile for each subject that will influence the PC scores. This shows the importance of the use of means and standard deviations of the PC scores in mortality prediction model.

Of course, things are more complicated once we start interpreting every component. Instead, in Table 3.1 we will provide just the interpretation of

Result	Interpretation	Surrogates
m_{i1}	Individuals with higher levels of overall activity during the day, and those who have higher early afternoon activity relative to early AM are associated with later mortality	Average TLAC
s_{i6}	Individuals who are more variable in the start time of their daily activity are associated with earlier mortality.	1.Standard deviation of ratio of mid-day to morning/afternoon activity. 2. Standard deviation of the difference in average activity during peaks/troughs highlighted by PC6.

Table 3.1: Interpretation of the results of fPCA

those components and summaries that were found to be predictive of the outcome.

3.4 Statistical Analysis

The demographic and clinical characteristics of the participants are presented in Table 3.2. They are separated by mortality status five years after the accelerometry study. For continuous variables, the mean is reported along with the standard deviation (in parentheses). For binary or categorical variables, the number of study participants in each category is reported along with the percent number of participants (in parentheses) out of the total number in the corresponding alive or dead category. Variables are ranked in decreasing order of their predictive performance as measured by the receiver operating characteristic curve (ROC) in single predictor logistic regression with the 5-

year all-cause mortality as response. The total activity count is the top-ranked individual 5-year mortality predictor (AUC = 0.771) while age is a close second (AUC = 0.758).

3.4.1 Mortality prediction models

Our main goals are to: (1) rank predictors in terms of their 5-year mortality predictive performance; and (2) identify the best subset of 5-year mortality predictors. To ensure that results are generalizable to the US population, weights were calculated for the selected subset of participants using the function **reweight_accel()** in the **rnhanesdata** package. After reweighting, we employed survey-weighted logistic regression using the function **svyglm()** in the **R** package **survey**. Variables are ranked according to the 10-fold complex survey weighted cross validated AUC in univariate models, where one predictor at a time is used to predict 5-year mortality. To select the best subset of 5-year mortality predictors, we use forward selection survey weighted logistic regression with the weighted cross-validated AUC as the optimization criterion. The variables in the final model were selected to maximize the cross-validated AUC, though we also report the Akaike's information criterion (AIC) [Lumley and Scott, 2015] and the efficient parsimony information criterion (EPIC) [Shinohara et al., 2011].

3.5 Results

Participant characteristics by mortality status are provided in Table 3.2. The mean age of the study sample was 65.9 (± 9.6, range 50.0-84.9) years. The proportions of men (51%) and women (49%) were similar with a larger pro-

portion of men (65% of mortalities) dying within 5 years of the follow-up. The participants who died within 5 years were on average 8.4 years older and had less time in MVPA, higher active to sedentary/sleep/non-wear transition probability ($\text{ASTP}_{\text{sl/nw}}$), TAC, and TLAC, lower sedentary/sleep/non-wear to active transition probability ($\text{SATP}_{\text{sl/nw}}$), and more sedentary/sleep/non-wear time. There was a larger proportion of nondrinkers and smaller proportion of moderate drinkers among the individuals who died compared to the group who did not. The proportions of smokers and former smokers were higher among the individuals who died. There was a larger proportion of individuals with less than high school education and a smaller proportion of individuals with more than high school education who died versus those who survived. The proportion of participants with CHF, coronary heart disease, and diabetes was higher among those who died within 5 years. There was a slightly larger proportion of deceased participants with underweight BMI, whereas the proportions of deceased and alive participants with normal and overweight BMI was similar. Finally, the proportion of alive individuals was slightly higher among Mexican Americans relative to other race categories, whereas the proportions of alive and deceased participants were similar in other race categories.

Table 3.2 shows the predictors' ranking according to AUC in univariate logistic regression models, where each mortality prediction model was fit with one predictor at a time. TAC is the strongest individual predictor of 5-year mortality (AUC = 0.771) with age (AUC = 0.758) and MVPA (AUC = 0.745) being close a second and third, respectively. The transition probability from active to sedentary/sleep/non-wear ($\text{ASTP}_{\text{sl/nw}}$, AUC = 0.733) and total sedentary time (sedentary/sleep/non-wear time, AUC = 0.728) round out the list of the top five individual predictors of 5-year all-cause mortality.

The next eight most predictive variables (excluding mobility difficulty) are all derived from accelerometry data with AUCs from 0.721 to 0.658. These results indicate that accelerometry-derived variables are strong predictors of mortality that outperform traditional risk factors including smoking, total cholesterol, gender, cancer, stroke, diabetes, and coronary heart disease.

		Alive	Dead	
Rank	Characteristics	Mean(SD)/N(%)[a]		AUC[b]
1	TAC	218013 (111831.2)	136362.7 (94487.5)	0.77
2	Age	65.1 (9.3)	73.5 (8.9)	0.757
3	MVPA	14.7 (17.3)	6.5 (12.1)	0.748
4	ASTP	0.29 (0.08)	0.37 (0.11)	0.734
5	Sedentary time	1102.5 (104.9)	1183.6 (110.6)	0.728
6	TLAC	2811.5 (704.6)	2281 (746.1)	0.722
7	TLAC 4PM-6PM	381.1 (112.1)	309.7 (116.6)	0.694
8	TLAC 12PM-2PM	410.2 (113.8)	333.5 (125.7)	0.692
9	TLAC 2PM-4PM	398.1 (116.6)	323 (121.6)	0.692
10	TLAC 6PM-8PM	320.7 (118.5)	250.8 (109.1)	0.691
11	TLAC 10AM-12PM	411.1 (127)	335.3 (132.3)	0.681
12	Mobility problem	766 (28.6%)	171 (58.2%)	0.672
13	SATP	0.08 (0.02)	0.07 (0.02)	0.66
14	SD on PC 6 (surrogate)	0.7 (0.27)	0.57 (0.25)	0.657
15	TLAC 8AM-10AM	344.7 (153)	282.3 (149.2)	0.63
16	Education			0.611
	Less than high school	819 (30.6%)	121 (41.2%)	
	High school	657 (24.6%)	79 (26.9%)	

	More than high school	1199 (44.8%)	94 (32%)	
17	TLAC 8PM-10PM	209 (122.5)	166 (104.4)	0.603
18	TLAC 6AM-8AM	170.8 (153.3)	127.7 (122.2)	0.594
19	Drinking Status			0.593
	Moderate Drinker	1341 (50.1%)	99 (33.7%)	
	Non-Drinker	1106 (41.3%)	158 (53.7%)	
	Heavy Drinker	153 (5.7%)	27 (9.2%)	
	Missing alcohol	75 (2.8%)	10 (3.4%)	
20	Smoking Status			0.574
	Never	1235 (46.2%)	90 (30.6%)	
	Former	1004 (37.5%)	137 (46.6%)	
	Current	436 (16.3%)	67 (22.8%)	
21	CHF	117 (4.4%)	49 (16.7%)	0.569
22	BMI			0.56
	Normal	666 (24.9%)	97 (33%)	
	Underweight	22 (0.8%)	7 (2.4%)	
	Overweight	1044 (39%)	100 (34%)	
	Obese	943 (35.3%)	90 (30.6%)	
23	Cancer	383 (14.3%)	73 (24.8%)	0.559
24	Diabetes	441 (16.5%)	74 (25.2%)	0.556
25	Gender			0.554
	Male	1324 (49.5%)	191 (65%)	
	Female	1351 (50.5%)	103 (35%)	
26	Stroke	131 (4.9%)	42 (14.3%)	0.548
27	CHD	195 (7.3%)	47 (16%)	0.548

28	TLAC 2AM-4AM	15.4 (52.5)	18.4 (45.1)	0.522
29	Race			0.514
	White	1553 (58.1%)	199 (67.7%)	
	Mexican American	500 (18.7%)	33 (11.2%)	
	Other Hispanic	53 (2%)	3 (1%)	
	Black	482 (18%)	52 (17.7%)	
	Other	87 (3.3%)	7 (2.4%)	
30	TLAC 12AM-2AM	25.1 (61.4)	24.7 (47.9)	0.509
31	TLAC 10PM-12AM	86.3 (100.5)	75.4 (80.9)	0.508
32	TLAC 4AM-6AM	39.2 (84.6)	34.5 (64.3)	0.504
33	Wear time	877.1 (134.5)	892.4 (171.4)	0.459

[1]

Table 3.2: Demographic and Clinical Characteristics Separated by Alive and Deceased Status 5 Years After Participation in the Accelerometry Study, National Health and Nutritional Examination Survey Pooled Cohorts Study, United States, 2003-2006.

We also consider multipredictor models and use forward selection that adds one variable at a time by maximizing the cross-validated AUC. Figure 3.4 displays the Akaike's information criterion, efficient parsimony information criterion, and AUC at each stage of the forward selection process. The scale

[1]BMI = body mass index; CHD = coronary heart disease; CHF = congestive heart failure; MVPA = moderate-to-vigorous physical activity; SATP = transition probabilities from sedentary to active; SD = standard deviation; TAC = total activity count; TLAC = total log activity count.

[a]: For continuous variables, the mean is reported with the standard deviation shown in parentheses. For binary or categorical variables, the number of study participants in that category is reported with the alive/deceased specific prevalence of each category in parentheses.

[b]:Variables are ranked by their individual predictive ability as measured by AUC in single predictor logistic regressions with 5-year all-cause mortality as the outcome.

for Akaike's information criterion and efficient parsimony information criterion is shown on the left y-axis, whereas the scale for AUC is shown on the right y-axis. The final 5-year mortality prediction model selected based on the cross-validated AUC criterion contains 13 predictors (arranged in the order of their selection): TAC, age, smoking status, CHF, drinking status, $\text{ASTP}_{\text{sl/nw}}$, mobility problem, gender, the surrogate for the SD on the sixth PC (SD on PC 6 surrogate), diabetes, education, TLAC 12-2 AM, and stroke.

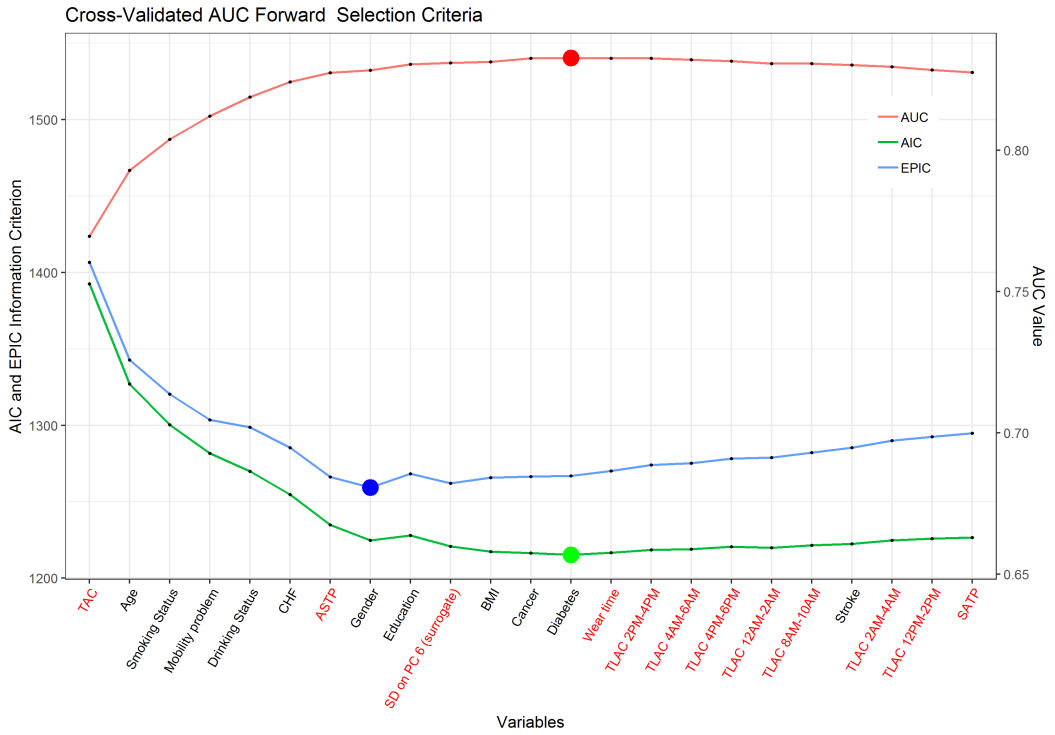


Figure 3.4: Model selection criteria plotted as a function of the variables added in the forward selection procedure. Predictors are shown on the x-axis, with accelerometry predictors in red. The AIC and EPIC information criterion values are shown on the left y-axis and the AUC values are shown on the right y-axis. This figures shows the best model for each of the three criteria at the colored dots. It also shows the change of AIC, EPIC and AUC as each variable is added into the model. Data source: National Health and Nutritional Examination Survey Pooled Cohorts Study, United States, 2003-2006.

Figure 3.5 displays the correlation plot between age and all activity-derived variables. Age has high negative correlations with TAC, TLAC, and positive correlation with sedentary time. TAC is highly correlated with most activity-derived measures, including MVPA, $\text{SATP}_{\text{sl/nw}}$, sedentary time, TLAC, and TLAC 4-6 PM, 6-8 PM, 2-4 PM, 12-2 PM, 10-12 PM, 8-10 AM, 6-8 AM.

$\text{ASTP}_{\text{sl/nw}}$, sedentary time, TLAC, and $\text{SATP}_{\text{sl/nw}}$ are the most highly correlated with multiple other variables. The surrogate for the standard deviation on the sixth PC (s_{i6}) has low correlation with other activity-derived variables, which may explain why it was selected in the joint model in addition to the other covariates.

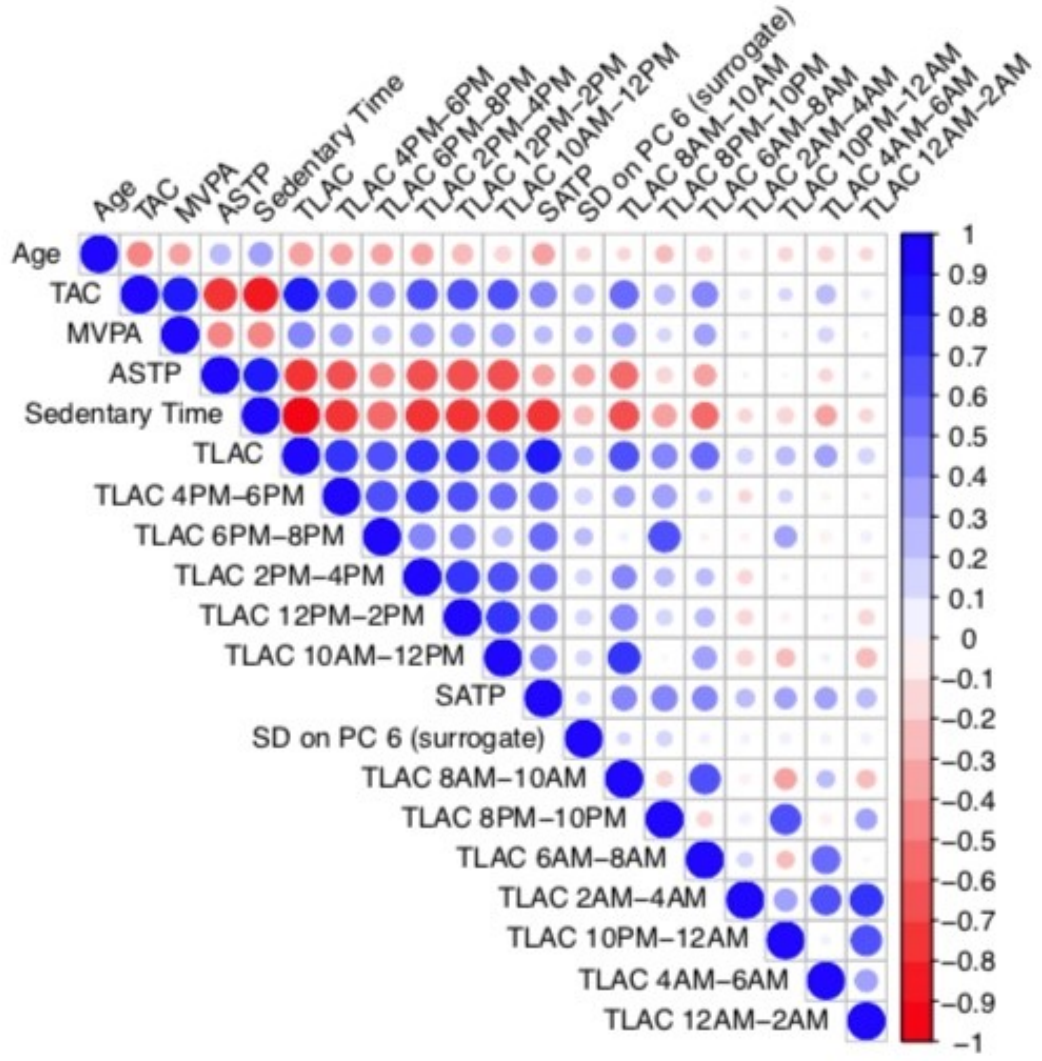


Figure 3.5: Correlation plot between age and accelerometry derived measures. National Health and Nutritional Examination Survey Pooled Cohorts Study, United States, 2003-2006.

Table 3.3 provides the results (point estimates odds ratio (OR), and confidence intervals (CIs), p-value) for the 13-variable model obtained via forward selection using the cross validated AUC, after accounting for all other predictors in the model. The mortality risk increases significantly with age (OR = 1.087, CI: (1.063, 1.112); p < .001) and history of coronary heart failure (OR = 2.175, CI: (1.177, 3.930); p = .013). Females have a lower probability of death (OR = 0.523, CI: (0.332, 0.817); p = .007), whereas current smokers (OR = 2.219, CI: (1.412, 3.478); p = .002) have a higher mortality risk than nonsmokers. Nondrinkers (OR = 1.759, CI: (1.165, 2.677); p = .010) and heavy drinkers (OR = 2.620, CI: (1.148, 5.673); p = .018) have higher 5-year mortality risk compared with individuals who consume alcohol moderately. When adjusted for age and other risk factors, including accelerometry-derived variables, the mortality risk was not statistically associated with higher total activity (OR = 1.007, CI: (0.508, 1.832); p = .982) but was positively associated with the active to sedentary transition probability ($ASTP_{sl/nw}$; OR = 1.465, CI: (1.078, 1.993); p = .016). Finally, higher values of the surrogate for the SD on the sixth PC (SD on PC6 surrogate; OR = 0.748, CI: (0.629, 0.885); p = .002) are associated with a lower probability of 5-year all-cause mortality. Although TAC is the most predictive individual variable of the 5-year mortality in single-regression models, its importance is substantially reduced after forward selection. This likely happens because many accelerometry-derived variables are highly correlated with age (Figure 3.5).

	Estimate	p-value	Confidence interval (95%)
Intercept	0.000	<.001	(0.000, 0.001)
Total activity count	1.007	.982	(0.508, 1.832)
Age	1.087	<.001	(1.063, 1.112)
Former smoker	1.394	.176	(0.835, 2.345)
Current smoker	2.219	.002	(1.412, 3.478)
Coronary heart failure: yes	2.175	.013	(1.177, 3.930)
Nondrinker	1.759	.010	(1.165, 2.677)
Heavy drinker	2.620	.018	(1.148, 5.673)
Missing alcohol	2.111	.106	(0.752, 5.193)
ASTP	1.465	.016	(1.078, 1.993)
Mobility problem	1.726	.028	(1.057, 2.816)
Gender: female	0.523	.007	(0.332, 0.817)
SD on PC 6 (surrogate)	0.748	.002	(0.629, 0.885)
Diabetes: yes	1.241	.310	(0.780, 1.937)
High school education	0.992	.973	(0.582, 1.694)
More than high school education	0.794	.309	(0.489, 1.294)
TLAC 12-2 AM	1.137	.099	(0.958, 1.322)
Stroke: yes	1.213	.505	(0.636, 2.227)

Table 3.3: Estimated Final Model Coefficients Odds Ratio (OR) with Corresponding Standard Errors and Significance Values in the Final Complex Survey Design Model, National Health and Nutritional Examination Survey Pooled Cohorts Study, United States, 2003-2006

To study the added prediction performance of accelerometry-derived PA

variables, we started with the optimal model using forward selection with PA and non-PA variables. This model had 13 variables, a cross validated AUC = 0.838 and is summarized in Table 3.3. From this model we constructed a model without PA variables by removing TAC, ASTPsl/nw, the surrogate for the standard deviation on the 6^{th} PC (SD on PC 6 surrogate), and TLAC 12AM-2AM. The resulting model had a cross validated AUC = 0.798 with the following non-PA variables: age, smoking status, CHF, drinking status, mobility problem, gender, Diabetes, Education, and Stroke. The improvement in the continuous net reclassification index (NRI) [Pencina et al., 2010], when comparing the 9-predictor (without PA covariates) and 13 predictor (with additional PA covariates) models was strongly statistically significant (p-value $<$ 0.001). This indicates that there is strong evidence against the null hypothesis of no improvement in reclassification when accelerometry-derived PA measures are allowed to enter the prediction model.

3.6 Discussion

One of our goals was to compare the individual predictive power of physical activity measures to traditional measures. However, some of the activity measures can be highly correlated with other variables. Thus, to make the predictive accuracy of these variables comparable, we examine the effect of each variable in single predictor regression models. Table 3.2 illustrates the strong predictive performance of objective PA measures derived from measurements collected by a hip-placed accelerometer. These predictors substantially outperform established predictors of mortality in single predictor regression models.

There are important limitations to the results in Table 3.2. Indeed, they are based on single variable regressions, which provide ranking of predictors of 5-year all-cause mortality if one can measure only one variable at a time. This is useful, but one is often interested in building risk scores based on combinations of variables. Indeed, one could argue that accelerometer-derived PA measurements may be so predictive because they are highly correlated with age. The practical implication would be that objective PA measurements might not be modifiable. For this reason, we have conducted forward selection with a rich pool of potential mortality predictors including the standard demographic, behavioral and comorbidities. Another limitation is the exclusion of interaction terms from the analysis. Unreported results indicate that most predictive interactions were between age and objective PA predictors. While some of the interactions were significant, they did not fundamentally change the results. Thus, to preserve simplicity, the focus here is on main effects prediction.

The association between PA and time to death using Cox models in the NHANES population has been investigated in several publications [Fishman et al., 2016a], [Di et al., 2017]. Here we focused on the 5-year mortality instead of time-to-death because: (1) the model and the results are easy to communicate; (2) potential problems with Cox model assumptions are avoided; and (3) it is one of the standard horizons for prediction modeling [Fishman et al., 2016b], [Choudhury et al., 2019]. We further considered excluding all participants who died within 1 and 2 years from the time when the survey was conducted to avoid potential reverse causation. When doing that, results remained qualitatively consistent, identifying the same core predictors of mortality. However, since the focus is on mortality prediction (and not

causal associations), results are presented without removing mortality data for the first years.

To study the robustness of finding, we have also conducted the same analyses in two age subgroups: (1) participants age 50 to 70 (1828 alive and 98 deceased within 5 years); and (2) participants age 70 and above (853 alive and 199 deceased within 5 years). In both age subgroups, 8 and 9 out of the top 10 mortality predictors were accelerometry derived PA summaries, respectively.

The selection of the exact collection of PA summaries derived from accelerometry that predict the 5-year all-cause mortality outcome can vary when data sets are slightly modified. This is likely due to the strong correlation among the objective PA summaries as well as to their correlation to other, established, risk factors. A better understanding of these relationships may further strengthen our understanding of the mutual effects of activity and other risk factors and their joint effect on mortality risk. However, these results indicate that: (1) there is a strong association between PA summaries derived from accelerometry and mortality; (2) these effects are encapsulated in different dimensions of PA measures; and (3) the combined effects of these summaries are independent of other, well known, mortality risk factors.

Given the large body of research on possible predictors of mortality, we conclude that PA summaries derived from accelerometry should become one of the top standard predictors of mortality risk. These measurements are becoming increasingly routine, are cheap and nonintrusive. Once they are normalized across cohorts and can be quantified in terms of easy to understand activities, duration, and timing, this could lead to more targeted PA intervention research.

Chapter 4

Microbiome Data Analysis

In this chapter, we will introduce microbiome data and the challenges in analyzing these data. Specifically, we will discuss the idea of filtering, an approach to detect and remove contaminant taxa. After a thorough literature review, we will present a novel filtering method by [Smirnova et al., 2018a] and my improvement to the efficiency of the filtering algorithm. The corresponding R package **PERFect** is then provided and two applications are presented to illustrate the robustness of the method.

4.1 Microbiome Data

The human microbiome is the collection of microbial organisms (microbiota) that live both inside and on the surface of humans [Matsen, 2014]. For example, they can be found on the skin, in the saliva, in the lung and especially in the gut and gastrointestinal tract. The five types of microorganisms that make up the human microbiome are bacteria, archaea, fungi, protozoans and viruses. Among them, bacteria are the most abundant members of the human micro-

biome: it has been estimated that there are ten times as many bacteria than there are human cells within each individual. Since the majority of the human microbiota are found in the gut, they significantly contribute to our immune system development, nutrition and drug metabolism [Maurice et al., 2013]. It has been shown that changes in microbiota composition play an important role in the development of multiple diseases including inflammatory bowel disease [Huttenhower et al., 2014], diabetes [Proctor, 2014, Pascale et al., 2019], preterm birth [Callahan et al., 2017, DiGiulio et al., 2015], and liver diseases [Puri et al., 2018]. Hence, studies of microbiota association with human disease states have received increasing attention over the last decade [Nguyen et al., 2015].

Next generation sequencing (NGS) of the 16S rRNA marker is currently among the most widely used methods for microbial organisms identification. In these studies, samples collected at different body sites (e.g., vaginal swab, stool or blood) give counts of DNA fragments which are then grouped into similar microbial organisms, usually referred to as taxa; in statistical terminology, these are random variables. In contrast to other measurements in genomics or metabolomics, microbiome data are very sparse as many taxa are rare and often have zero counts in most samples. Hence, the resulting data, usually referred to as the 'taxa table' are typically high dimensional.

4.1.1 Challenge of Microbiome Data

The extreme levels of sparsity in microbiome datasets is one of the major challenges in data analysis. Indeed, it is not unusual to have over 90% of the counts being zero in these data as they contain a large number of rare taxa observed

in as few as 1 to 5% of samples. However, recent microbiome quality control studies show that the majority of rare taxa are caused by contamination and/or sequencing errors. Potential sources of contamination are bacteria that are frequently handled in the lab, those that reside on the skin of lab workers, or in the extraction kits [Salter et al., 2014]. Several studies have been conducted using 'mock' samples curated so that they consist of known microbial species in prescribed proportions and, after cultivation, the samples are sequenced using NGS technology to identify the taxa and evaluate the effects of such contamination on the observed taxa counts [Brooks et al., 2015]. Errors, especially due to misclassification, arise as the sequencing technology employs a combination of statistical and computational algorithms that make assumptions about identifying nucleotide bases [Cacho et al., 2015] and for assembling the DNA fragments during the alignment process [Li and Homer, 2010]. Overall, contamination and sequencing errors lead to either falsely identifying taxa that were not in the sample or misclassifying the taxa of DNA fragment reads. The most common approach to address this problem is filtering, or removing spurious taxa from the 16S data set, which is a variation of an ad hoc, albeit simple, procedure. For example, one of the most widely used techniques for filtering in microbiome studies selects taxa that have a number of counts above $m = 0$ in at least n samples.

In practice, it is often of interest to use taxa as covariates to predict disease outcomes and understand their association with the host health. Examples include predicting small intestine bacteria overgrowth (SIBO) condition using taxa sequenced from the intestine [Leite et al., 2019], testing whether dietary interventions shape gut microbiota [Albenberg et al., 2012] and understanding the impact of a probiotic intervention on the composition of the

human microbiota [Lahti et al., 2013]. However, in high dimensional setting (large number of variables), it is challenging to find a few important predictors [Fan and Lv, 2008]. Indeed, with the high dimensionality p, computational cost and prediction accuracy are two top concerns for any statistical procedure, especially in the presence of extreme sparsity. Hence, dimension reduction for sparse data is often recommended to reduce computational burden by effectively identifying the subset of important predictors and improve estimation accuracy by using well-developed lower dimensional methods.

In microbiome setting, contaminant and rare taxa may be considered as unimportant predictors. While several techniques have been proposed to detect and remove them, the literature in this research area is scarce. One approach, developed by [Knights et al., 2011] and implemented in the R package **sourcetracker**, relies on microbial source tracking to identify the proportion of contaminant taxa in each sample by matching the taxa table to the database of known contaminants.

However, this method does not detect individual contaminant taxa that should be removed from the data set. [Davis et al., 2018] addressed this problem by introducing the **decontam** R package that identifies contaminants by: (1) inversely correlating taxa frequencies with sample DNA concentrations; and (2) using the prevalence of sequenced negative controls [Salter et al., 2014]. A major practical limitation of this method is that the auxiliary data from DNA quantitation, which is in most cases intrinsic to sample preparation or negative controls data, might not be available.

Recently, [Smirnova et al., 2018a] introduced a filtering loss measure and a principled filtering test, PERFect, for deciding which taxa to remove. In contrast to the standard procedures, which assume that taxa in a biological

network are isolated, PERFect filters out taxa with insignificant contribution to the total covariance. This method relies on ranking taxa importance, measuring their contribution to the total covariance, and quantifying the chance that the loss increase for a set of filtered taxa is due to randomness. The two principled filtering methods, simultaneous and permutation algorithm, rely on estimating the null distribution for the increase in filtering loss due to each taxon. The simultaneous approach fits one distribution for the filtering loss differences for all taxa, whereas the permutation approach generates a distribution containing k permutations of filtering loss differences and fits it for each taxon. Thus, one major limitation of our initial software implementation was the computational intensity of the PERFect permutation method, which was shown to be both a statistically rigorous and highly effective filtering approach. Here, I introduce the fast implementation of the permutation PERFect method that efficiently selects a small subset of taxa to build the distribution necessary to assess the taxon's significance. The process of selecting this taxa subset is performed using an unbalanced binary search algorithm [Morin, 2013] that optimally finds the set of taxa to be removed without building the permutation distribution and computes the p-values for all taxa. The proposed approach successfully reduces algorithm running time by almost four times.

The effects of filtering are further evaluated on two major exploratory analyses used in microbiome research: alpha and beta diversity. The methods were applied to two data sets, namely the MicroBiome Quality Control (MBQC) project from [Sinha et al., 2015] and the laboratory contamination dataset (Salter) from [Salter et al., 2014]. Results show that the filtering methods reduce the magnitude of differences in alpha diversity for samples containing the same bacteria processed at different MBQC project labs. Filtering further

reduces dissimilarity between samples (beta diversity) that contain the same microbiome and potentially alleviates technical variability. In the next section, we will be introduced to the setup for MBQC data, which will be used as a guided example, to reinforce our understanding throughout the whole methods section.

4.1.2 The MicroBiome Quality Control data

Consider the dataset from the MBQC project, a collaborative effort designed to comprehensively evaluate sample processing and computational methods for human microbiome data analysis [Sinha et al., 2015]. There are four types of samples in this dataset: (1) 11 unique fresh stool samples; (2) seven unique freeze-dried stool samples; (3) two unique chemostat samples generated from a Robogut; and (4) two artificial colonies representing the gut and oral cavity. These samples were randomly sequenced at 15 laboratories and then randomly distributed to 9 bioinformatics facilities for microbiome analyses. Here, we consider the oral artificial communities data comprised of 22 true taxa from the human oral cavity. The MBQC data identified a total of $27,140$ taxa across the four types of samples. For the purposes of this analysis, $14,861$ taxa that have a 0 count across all oral artificial community samples are excluded; 1277 taxa that match names at the species level are combined; finally, $10,210$ taxa that appeared in less than 5% of the samples are removed. The final dataset considered for this analysis contains 1016 samples and 792 taxa. A limitation of this dataset is that the samples were created from the species in prescribed proportions; however, after the samples were processed many taxa were only identified up to the genus level (higher order phylogenetic hierarchy). As

a consequence, only two signal taxa, Veillonellaceae Veillonella Parvula and Coriobacteriaceae Eggerthella Lenta, were correctly detected while the other 20 signal species corresponded to one of the 184 taxa identified at the genus level.

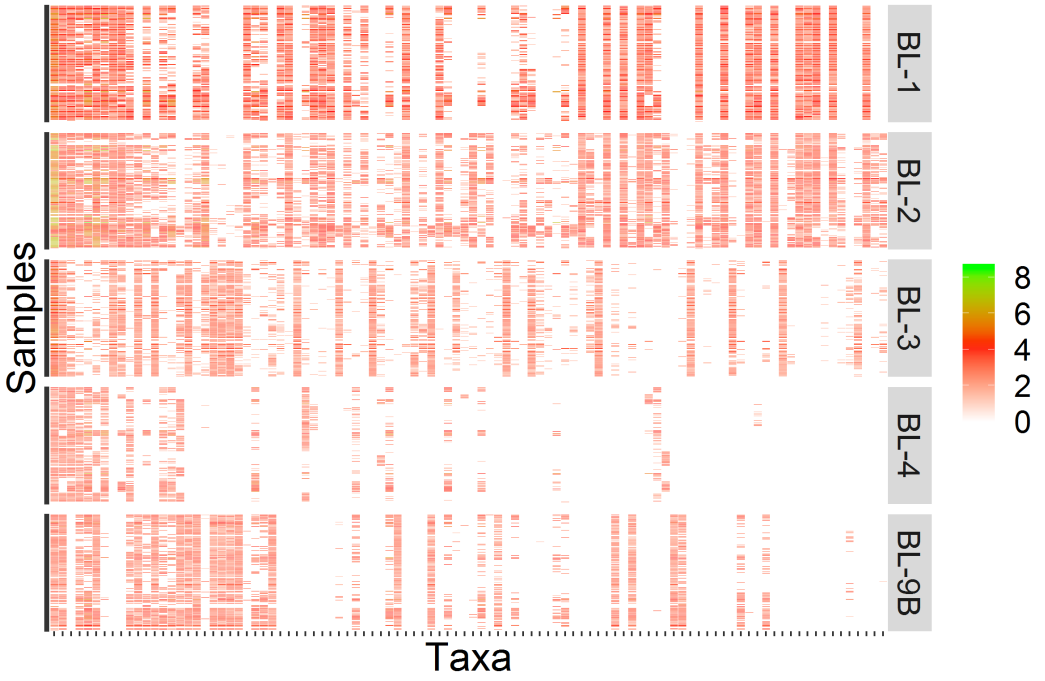


Figure 4.1: The heatmap of 100 observed taxa on the log-scale, with taxa on the x-axis arranged in decreasing abundance order and samples on the y-axis arranged by processing institutes. Source: [Sinha et al., 2015]

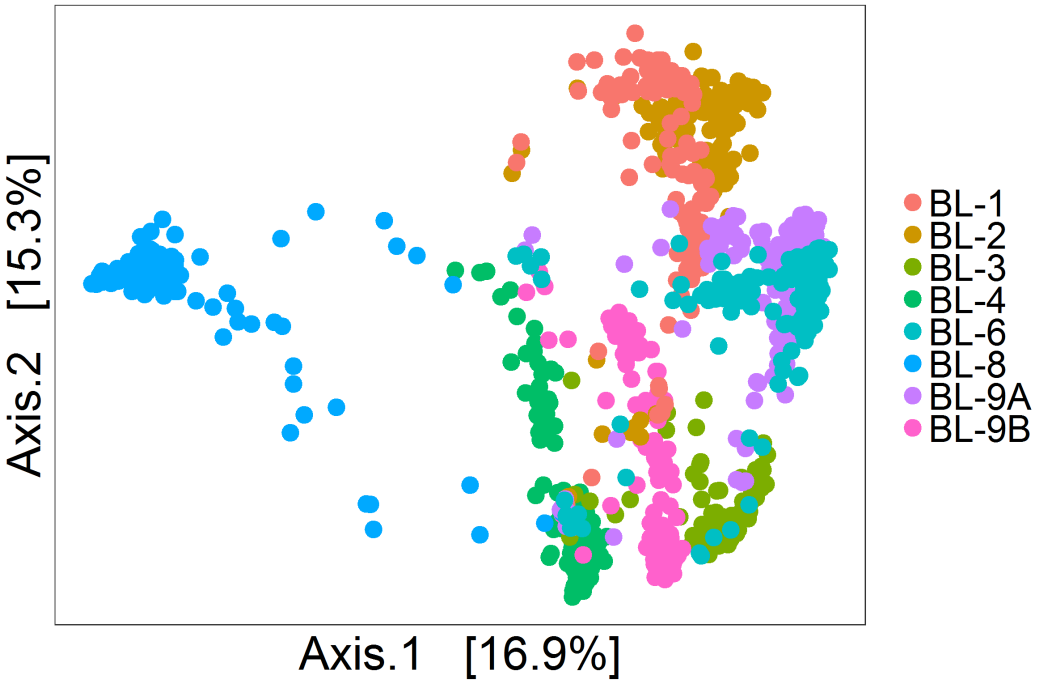


Figure 4.2: The multidimensional scaling plot of 1016 samples, colored by the processing institutes. Source: [Sinha et al., 2015]

Figure 4.1 displays the log-counts heat map for the 100 most abundant taxa, arranged in decreasing order of abundance. The lighter-colored areas of the heatmap in the lower right corner indicate unobserved taxa, showing the decrease of signal strength with different processing institutes/labs. Figure 4.2 displays the Bray-Curtis distance [Quaak and Kuiper, 2011] multidimensional scaling (MDS) plots for 1016 samples from the heat map on the left. The first two principal components (PCs) that explain 32.2% of variability are shown on the plot. The distance between samples varies as the labs change, indicating that the samples processed at different institutes appear to have dramatic differences on MDS plots even though they contain exactly the same signal species. These differences are likely due to the bioinformatics pipeline

that varies between different labs. This problem motivates filtering, which removes rare taxa displayed in columns on the right-hand-side of the heatmap in Figure 4.1. Left unresolved, this problem may cause a number of practical issues including: (1) falsely inflating within sample diversity, called α-diversity [Park and Allaby, 2017a]; (2) obscuring true distances between samples, called β-diversity [Park and Allaby, 2017b]; and (3) interpreting rare taxa as disease biomarkers (especially in low sample biomass environments).

4.1.3 Methodology

Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ be a taxa counts table, where each column $\boldsymbol{x}_j \in \mathbb{R}^n$, $j = 1, \ldots, p$ contains the j^{th} taxon counts observed across n samples. Filtering removes columns of the taxa table, X, that correspond to the subset of taxa $J \subset \{1, \ldots, p\}$ according to a particular criterion. For example, if X contains $p = 20$ taxa columns, labeled $T_1, \ldots, T_{20}$ and taxa $\{T_4, T_5, T_{15}\}$ (in the $4^{th}, 5^{th}$ and 15^{th} columns of X) do not contribute to the signal, then $J = \{4, 5, 15\}$. [Smirnova et al., 2018a] introduced *simultaneous* and *permutation* filtering approaches and implemented them in the **PERFect** package. Here these approaches are briefly described.

PERFect derives a filtering threshold based on measuring the loss of taxa contribution to the total covariance of the data. Specifically, the filtering loss due to removing a group of taxa J is defined as

$$FL(J) = 1 - \frac{\|X_{-J}^T X_{-J}\|_F^2}{\|X^T X\|_F^2}, \tag{4.1}$$

where X_{-J} is the $n \times (p - |J|)$ dimensional matrix obtained by removing the columns indexed by the set J from the data matrix X, $||\cdot||_F^2$ is the Frobenius

norm, and $|J|$ is cardinality of the set J (here, the number of removed taxa). The filtering loss $FL(J)$ is a number between 0 and 1, with values close to 0 if the set of taxa J has small contribution to the total covariance and 1 otherwise.

To compute the filtering loss, taxa are first re-arranged in ascending order of importance, based on the assumption of taxa contribution to the signal. The most common filtering criterion used in practice is based on taxa abundance. This is equivalent to arranging taxa in ascending order from left to right according to their number of presences in the n samples, defined for the j^{th} taxon as

$$NP(j) = \sum_{i=1}^{n} I(x_{ij} > 0), \tag{4.2}$$

where x_{ij} is the i^{th} element in $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$ from the j^{th} column of X, and $I(\cdot)$ is the indicator function. For example, if the taxon T_1 appears in only 1 sample, it is the least abundant taxon and will show up to the left of the table, whereas if the taxon T_2 appears in every sample, it is the most abundant taxon and will be rearranged to the right of the table. Once taxa are ordered, the filtering loss is calculated sequentially by removing taxa from left to right. The statistical threshold for separating the signal from noise taxa is defined based on the location of dramatic increases in the filtering loss. This statistic is the key ingredient for deciding which increases in filtering loss can be attributed to randomness and which increases correspond to true signal in the data. It is defined for the $(j+1)^{th}$ taxon as

$$DFL(j+1) = FL(J_{j+1}) - FL(J_j), \tag{4.3}$$

where $FL(J_j)$ is the filtering loss from removing the first j taxa and $FL(J_{j+1})$ is the filtering loss from removing the first $j+1$ taxa. Assuming that $P = \{p_{ij}\}_{n\times p}$ is a matrix of the true relative abundance of microbe j in sample i ($\sum_j p_{ij} = 1$ for each sample i), the theoretical quantity for the filtering loss when a group of taxa J is removed is defined as

$$\mathcal{F}_J = 1 - \frac{||P_{-J}^T P_{-J}||_F^2}{||P^T P||_F^2},$$

which is equal to 0 if the group of taxa J is included erroneously. Then $d\mathcal{F}_{j+1} = \mathcal{F}_{J_{j+1}} - \mathcal{F}_{J_j}$ is the theoretical improvement to the signal from adding the taxon $j+1$. $\mathcal{F}_J$ and $d\mathcal{F}_{j+1}$ are estimated using the filtering loss $FL(J)$ (4.1) and corresponding differences in filtering loss $DFL(j+1)$ (4.3) statistics. Therefore, for each taxon, we test

$$H_0 : d\mathcal{F}_{j+1} = 0 \qquad \text{vs.} \qquad H_A : d\mathcal{F}_{j+1} > 0.$$

In the following two sections, we will introduce the main functions of the **PERFect** package and describe two approaches to building the distribution of $d\mathcal{F}$ using $DFL(j+1)$ values to distinguish noise and true signal in the data.

4.1.4 The PERFect Package

The main input for all functions in the **PERFect** package is a taxa table of class matrix or data frame. The major functions provided in this package are listed in Table 4.1.

Function	Description
`FL_J`	Calculate filtering loss due to removing a group of J taxa.
`FiltLoss`	Calculate filtering loss sequentially for removing a set of J_j taxa for $j = 1, ..., p$.
`DiffFiltLoss`	Calculate differences in filtering loss due to removing a set of J taxa sequentially.
`PERFect_sim`	Perform the simultaneous filtering test.
`PERFect_perm`	Perform the permutation filtering test.
`pvals_Plots`	Plot all the p-values from the results of a filtering test.

Table 4.1: Major functions in the package **PERFect**

4.2 Filtering algorithms

4.2.1 Simultaneous filtering

This method assumes that a large percentage of the taxa has low signal ($d\mathcal{F}_{j+1} = 0$) and the differences in filtering loss for p taxa are fit using **one** distribution. Due to the extreme left-skewed nature of the DFL measures, a log-transformation to the DFL data is required. The left part of the distribution of these log differences can be fit by quantile matching a Skew-Normal distribution [Azzalini, 2005] with location parameter ξ, scale parameter ω^2 and shape parameter α denoted by $\text{SN}(\xi, \omega^2, \alpha)$, while the right tail of the distribution is assumed to be generated by an unspecified distribution. The significance p-value of the set of the first j taxa J_j is calculated as:

$$p_j = P[X > \log\{DFL(j+1)\}], \tag{4.4}$$

where the random variable $X \sim \text{SN}(\widehat{\xi}, \widehat{\omega}^2, \widehat{\alpha})$ with parameters estimated from the log-transformed DFL data, and $\log\{DFL(j+1)\}$ is the log-transformed value of the filtering loss difference due to removing the J_j taxa. The simultaneous procedure is performed by the function `PERFect_sim()` in the package and its algorithm can be summarized in Algorithm 1, as given.

Algorithm 1 PERFect: simultaneous filtering

`Input:` Taxa table X, test critical value α
`Output:` Filtered OTU table X, p-value for each taxon
1: Order columns of X such that $NP(1) \leq NP(2) \leq NP(p)$
2: **for** `taxon j = 1, ..., p-1` **do**
Calculate $DFL(j+1)$ using (4.3) for $J_j = \{1, \ldots, j\}$
end
3: Using quantile matching fit the Skew-Normal distribution to the logarithm of the sample $DFL(j+1), j = 1, \ldots, p-1$ to obtain the null distribution $X \sim \text{SN}(\widehat{\xi}, \widehat{\omega}^2, \widehat{\alpha})$
4: Calculate the p-value p_{j+1} for $DFL(j+1), j = 1, \ldots, p-1$ as $p_{j+1} = P[X > \log\{DFL(j+1)\}]$
5: Average p-values for each set of 3 subsequent taxa
6: Filter out the set of taxa J_j with the first p-value such that $p_{j+1} \leq \alpha$

For example, the microbiome dataset from [Sinha et al., 2015] consists of 1016 samples and 792 taxa. The simultaneous filtering method removes 617 taxa, leaving a smaller dataset of 1016 samples and 175 taxa for further analysis. By default, the function `PERFect_sim()` takes the data table, X, as a matrix or data frame, orders it by the taxa abundance defined in (4.2), uses 10%, 25% and 50% quantiles for matching the log of DFL to a Skew-Normal distribution and then calculates the p-value for each taxon at the significance level of $\alpha = 0.1$. The function `PERFect_sim()` only needs a taxa table as the input, and other parameters are set to default.

```
> dim(Counts.mat)
[1]  1016  792
> res1 <- PERFect_sim(X = Counts.mat, Order = "NP",
+        quant = c(0.1,0.25,0.5), distr = "sn", alpha = 0.1)
> dim(res1$filtX)
[1]  1016  175
```

The plot of differences in filtering loss values by taxa and the histogram of log differences in filtering loss are extracted from the object `res1` above using the commands `res1$hist` and `res1$pDFL` and displayed in Figure 4.3. On the left panel, the spike of differences in filtering loss indicates the true signal taxa,

suggesting that taxa to the left of this spike should be removed from the data. The right panel shows a Skew-Normal distribution, which fits reasonably well the left part of the log differences in filtering loss using the quantile matching method. Although the robustness of this fit will vary based on the choice of quantiles, I suggest using $10\%, 25\%$ and 50% quantiles for matching since this method assumes that at least 50% of the taxa are not informative.

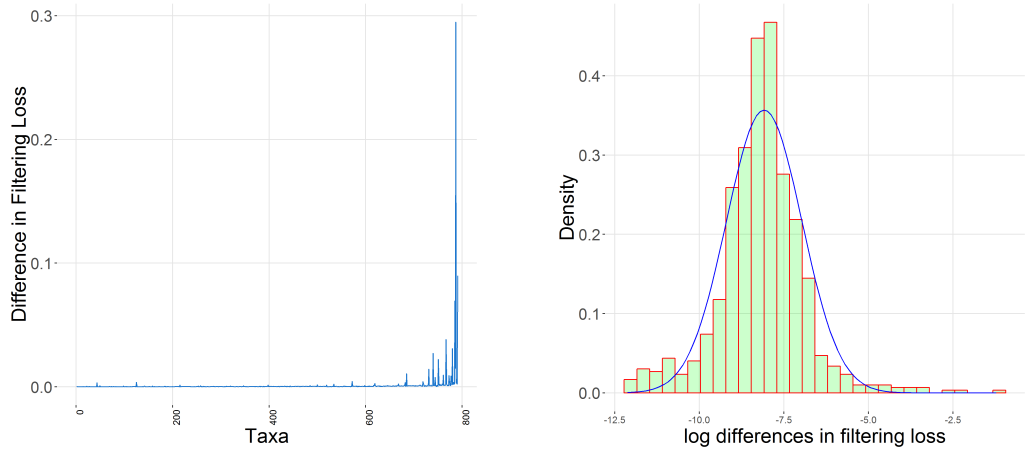


Figure 4.3: Left panel: plot of the differences in filtering loss for the taxa that are arranged in the order of increasing values of NP. Right panel: histogram of the log differences in filtering loss. The blue line indicates a $SN(\hat{\xi} = -8.12, \hat{\omega} = 1.12, \hat{\alpha} = 0.04)$ density fitted to the log-transformed data using 10%, 25% and 50% quantiles for matching. Data source: [Sinha et al., 2015].

4.2.2 Permutation filtering

The major difference between permutation simultaneous filtering is that instead of fitting the differences in filtering loss for p taxa using **one** distribution, we assume that at each step $(j+1)$, the corresponding $DFL(j+1)$ value

has its own Skew-Normal distribution. In order to estimate the parameters for this distribution, we randomly permute the labels of the taxa, calculate $DFL^*(j+1)$ for every permutation and fit a Skew-Normal distribution on these permuted DFL values. This approach ensures that a taxon with weak signal remains unimportant to any combination of other $(j+1)$ taxa. The remainder of this method is similar to the simultaneous method and its algorithm is summarized in Algorithm 2, as given.

Algorithm 2 PERFect: permutation filtering

`Input:` OTU table X, test critical value α
`Output:` Filtered OTU table X, p-value for each taxon
1: Run simultaneous PERFect algorithm to obtain taxa p-values $p_j, j = 1, \ldots, p$
2: Order columns of X such that $p_1 \geq p_2 \geq p_p$
3: **for** `taxon j = 1, ..., p-1` **do**
    Let $J_j = \{1, \ldots, j\}$
    Calculate $DFL(j+1)$ using (4.3)
4: **for** `permutation 1, ..., k` **do**
        Randomly select $J^*_{j+1} \subset \{1, \ldots, p\}$ with $|J^*_{j+1}| = j+1$
        Calculate $DFL^*(j+1)$ using (4.3)
    end
5: Using quantile matching fit the normal distribution to the logarithm of the sample $DFL^*(j+1), j = 1, \ldots, p-1$ to obtain the null distribution $X_{j+1} \sim \text{SN}(\widehat{\xi}_{j+1}, \widehat{\omega}^2_{j+1}, \widehat{\alpha}_{j+1})$
6: Calculate the p-value p_{j+1} for $DFL(j+1), j = 1, \ldots, p-1$ as $p_{j+1} = P[X_{j+1} > \log\{DFL(j+1)\}]$
end
7: Average p-values for each set of 3 subsequent taxa
8: Filter the set of taxa J_j with the first p-value such that $p_{j+1} \leq \alpha$

This procedure is performed by the function `PERFect_perm()`, which takes similar arguments as the function `PERFect_sim()`. For the same dataset from [Sinha et al., 2015], we applied the permutation filtering method using the decreasing order of p-values from the simultaneous filtering method above, resulting a dataset that retains 233 taxa and removes 559 taxa in total.

```
> res2 <- PERFect_perm(X = Counts.mat, Order = "pvals",
+         pvals_sim = res1, algorithm = "full", k = 10000,
+         quant = c(0.1,0.25,0.5), distr = "sn", alpha = 0.1)
```

```
> dim(res2$filtX)
[1]  1016  233
```

4.2.3 Fast permutation filtering

One drawback of the permutation filtering method is that it might be computationally expensive. Indeed, given that k permutations are performed for each taxon $j = 2, 3, \ldots, p$, the algorithm requires a total of $k(p-1)$ permutations, where k and p are large. Thus, I employ parallel processing and an unbalanced binary search algorithm [Morin, 2013] that *optimally finds the cut-off taxon j to remove the set of J_j taxa without building the permutation distribution and computing the p-values for all $p-1$ taxa.* Recall that the main goal is to find the first taxon j in the *ordered set of taxa* for which the filtering loss difference increases significantly. Thus, if the DFL value for a set J_j with, for example 10 taxa, is not significant, then the DFL values of the first 9 taxa in this set will not be significant as well and we can proceed to test the next set. However, there are two problems we need to consider: 1) we expect that the differences in filtering loss shown in Figure 4.3 (left panel) increase with j, thus the length of the search intervals need to be optimized to perform the least number of tests; 2) given taxa are ordered by a chosen order, if any taxon in the set J_j has a DFL value larger than that of the j^{th} taxon we tested (the last taxon of the ordered interval), then we need to test its significance.

We address the first problem by creating M sets $S_1, S_2, \ldots, S_M$ (shown in Figure 4.4) such that

$$S_m = \{T_1, T_2, \ldots, T_{\sum_{k=1}^{m} M-k+1}\}, \tag{4.5}$$

where T_m is the m^{th} taxon, $\sum_{k=1}^{m}(M-k+1)$ is the index of the last taxon for the subset S_m and $1 \leq m \leq M << p$.

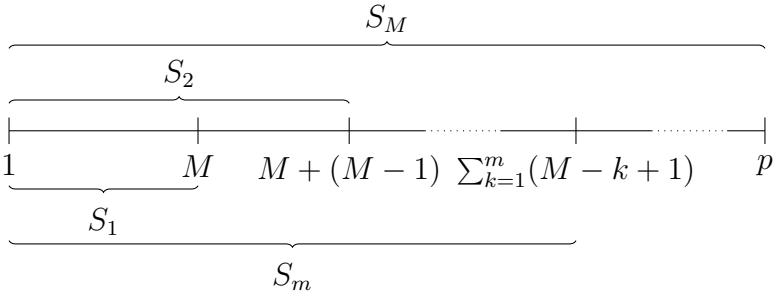


Figure 4.4: Taxa intervals tested by the fast permutation filtering algorithm.

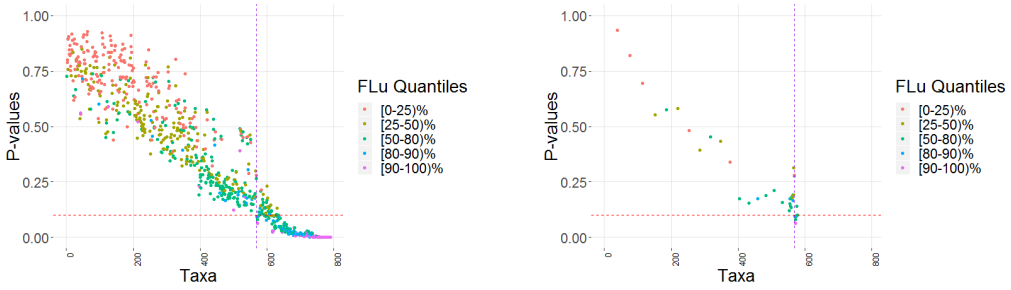
Thus, at every step we include smaller numbers of additional taxa as m increases (i.e. $M, M-1, \ldots, 1$). This approach minimizes the number of taxa distributions we need to calculate as taxa importance increases. We then we apply steps 3 to 6 of Algorithm 2, where in step 3 we set the taxa index $j \in \{\sum_{k=1}^{m} M-k+1\}_{m=1}^{M}$ and test whether the set $\{S_m\}_{m=1}^{M}$ should be removed until we find the interval with the first significant p-value. If set S_m is significant, it means that any taxon added to the previously tested set S_{m-1} might be significant. Therefore, we apply steps 3 to 6 of Algorithm 2 to each taxon added to S_{m-1} until the potential cut-off taxon, $T^{*}_{cut-off}$, with first significant p-value is identified. Finally, to address the second problem, we further test significance of all taxa with higher DFL values than the values of potential cut-off taxon. The first taxon with the smaller index that is significant is chosen as the final cut-off taxon $T_{cut-off}$. By default, this computation is given by `algorithm = "fast"`, but the user can modify it to `algorithm = "full"`, which will compute p-values for all taxa. The code for this result is shown below.

```
> system.time(res2 <- PERFect_perm(X = Counts.mat,
+                   Order = "NP", k = 10000,
+                   algorithm = "full",
+                   quant = c(0.1,0.25,0.5),
+                   distr = "sn", alpha = 0.1))
user       system       elapsed
573.18     0.37         775.87
> system.time(res3 <- PERFect_perm(X = Counts.mat,
+                   Order = "NP", k = 10000,
+                   algorithm = "fast",
+                   quant = c(0.1,0.25,0.5),
+                   distr = "sn", alpha = 0.1))
user       system       elapsed
5.11       2.00         236.16
```

Figure 4.5 illustrates the plot of permutation PERFect p-values calculated by the full and fast algorithm for the [Sinha et al., 2015] data. Although both methods achieve a similar cutoff taxon, the fast algorithm only calculates 43 out of 792 p-values, and hence is more computationally efficient.

4.3 Application and Evaluation

[Smirnova et al., 2018a] validated PERFect simultaneous and permutation filtering approaches using three mock community data sets ([Knights et al., 2011], [Ravel et al., 2011], and [Fettweis et al., 2012]), using the number of contaminant taxa correctly removed as an efficiency criterion. Here, I concentrate on the effects of filtering on downstream analyses, using the the two major



(a) Full algorithm (b) Fast algorithm

Figure 4.5: Left panel: Permutation PERFect p-values for every taxon from the [Sinha et al., 2015] data. Taxa on the x-axis, arranged in order of their abundance, are represented by points colored according to their individual filtering loss quantile values. The dashed horizonal red line indicates the $\alpha = 0.10$ cutoff. Taxa to the left of the dashed purple vertical line correspond to the set of filtered out taxa J and to the right of this line correspond to the set $\{-J\}$ of retained taxa. Right panel: Permutation PERFect p-values for 43 taxa (fast algorithm) from the [Sinha et al., 2015] data.

exploratory analyses used in microbiome research: alpha and beta diversity.

4.3.1 The MicroBiome Quality Control data

One of the main goals of the MBQC project is to understand major differences in technology and methods for analyzing human microbial data. This can be achieved by analyzing the observed taxa variation between handling lab and bioinformatics processing protocols. Here, we concentrate on the effect of bioinformatics processing laboratories on the observed oral mock community data measured by alpha and beta diversity, two of the most commonly used summaries in microbiome research. The left panel in Figure 4.6 shows the

logarithm of the Chao1 index (a measure of richness in alpha diversity) in the unfiltered and filtered data, colored by the processing institutes. Both filtering methods significantly reduce the richness in each dataset, especially the simultaneous method, but the overall variation pattern within labs remains unchanged. For example, samples coming from labs $BL-2$, $BL-6$ and $BL-9A$ consistently have the most variation in the index for every dataset. The right panel of Figure 4.6 shows the Shannon index, the most widely-used diversity metric which weights the number of species by relative evenness data [Reese and Dunn, 2018], in the unfiltered and filtered data. This plot and the summary statistics in Table 4.2 indicate a decrease in the Shannon index between the unfiltered and filtered data, which implies a reduction in the diversity and evenness of taxa. This phenomenon is expected since filtering highlights the signal contribution, causing a less even distribution of taxa in the data due to a large number of rarely observed variables.

		BL-1	BL-2	BL-3	BL-4	BL-6	BL-8	BL-9A	BL-9B
Median	Unfiltered	5.301	5.293	5.700	5.622	5.324	5.270	5.317	5.536
	Simultaneous	4.122	4.147	4.061	3.998	4.381	3.247	4.332	4.061
	Permutation	4.287	4.301	4.261	4.300	4.552	3.519	4.459	4.269
IQR	Unfiltered	0.137	0.175	0.165	0.446	0.181	0.062	0.161	0.477
	Simultaneous	0.135	0.231	0.123	0.420	0.205	0.456	0.280	0.083
	Permutation	0.166	0.235	0.127	0.475	0.224	0.489	0.288	0.098

Table 4.2: Summary statistics of the Shannon index for each processing lab.

In order to study the effect of filtering on differences across bioinformatics processing labs, we applied Dunn's test with a Benjamini-Hochberg correction for multiple testing to all possible pairwise Shannon alpha diversity comparisons between processing labs. Results are summarized in Table 4.3. Since

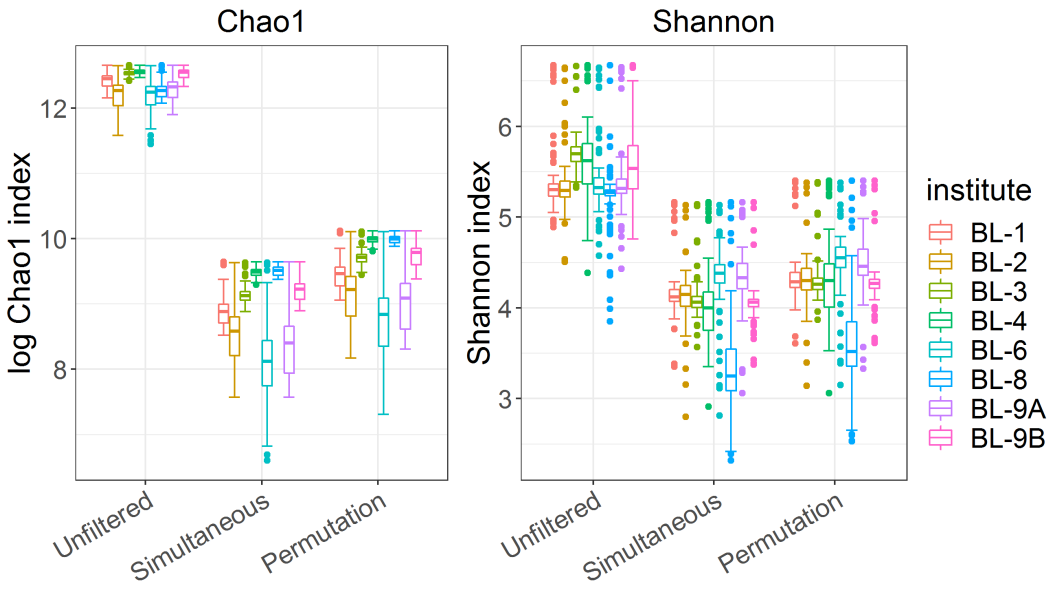


Figure 4.6: Left panel: The logarithm of the Chao1 index for the original data and two filtered data, colored by the dilution levels. Right panel: The Shannon index for the original data and two filtered data, colored by the dilution levels. Data source: [Sinha et al., 2015].

all samples contained the same mock communities, in the absence of technical variability, none of the differences should be significant. For the unfiltered data, 21 out of 28 possible pairs have significant differences in alpha diversity at the 0.05 significance level. Applying simultaneous and permutation filtering decreases differences in alpha diversity for most pairs. Moreover, *there are a total of* 4 *and* 8 *pairwise comparisons that are no longer significant at the* 0.05 *level after simultaneous and permutation filtering results were applied respectively.* While filtering does not remove all differences due to processing labs, these results indicate that it dramatically alleviates differences in alpha diversity estimates caused by lab-to-lab variability.

In order to study the effect of filtering on beta diversity, we calculated

	Unfiltered		Simultaneous		Permutation	
Comparison	Difference	P-values	Difference	P-values	Difference	P-values
BL-1 - BL-2	0.00	0.4990	0.20	0.4214	-0.06	0.4778
BL-1 - BL-3	-10.75	< 0.0001	2.52	0.0074	1.27	0.1307
BL-2 - BL-3	-10.83	< 0.0001	2.35	0.0115	1.33	0.1283
BL-1 - BL-4	-7.22	< 0.0001	3.90	0.0001	0.29	0.4173
BL-2 - BL-4	-7.28	< 0.0001	3.73	0.0001	0.35	0.4088
BL-3 - BL-4	3.91	0.0001	1.25	0.1243	-1.01	0.1816
BL-1 - BL-6	-1.63	0.0632	-6.35	< 0.0001	-7.10	< 0.0001
BL-2 - BL-6	-1.64	0.0646	-6.60	< 0.0001	-7.10	< 0.0001
BL-3 - BL-6	9.36	< 0.0001	-8.78	< 0.0001	-8.23	< 0.0001
BL-4 - BL-6	5.70	< 0.0001	-10.47	< 0.0001	-7.55	< 0.0001
BL-1 - BL-8	2.47	0.0090	11.30	< 0.0001	9.99	< 0.0001
BL-2 - BL-8	2.49	0.0089	11.19	¡ 0.0001	10.13	< 0.0001
BL-3 - BL-8	13.27	< 0.0001	8.51	< 0.0001	8.50	< 0.0001
BL-4 - BL-8	9.81	< 0.0001	7.60	< 0.0001	9.92	< 0.0001
BL-6 - BL-8	4.16	< 0.0001	17.93	< 0.0001	17.36	< 0.0001
BL-1 - BL-9A	-1.09	0.1535	-5.46	< 0.0001	-5.74	< 0.0001
BL-2 - BL-9A	-1.10	0.1583	-5.70	< 0.0001	-5.73	< 0.0001
BL-3 - BL-9A	9.66	< 0.0001	-7.86	< 0.0001	-6.88	< 0.0001
BL-4 - BL-9A	6.09	< 0.0001	-9.46	< 0.0001	-6.14	< 0.0001
BL-6 - BL-9A	0.51	0.3167	0.77	0.2455	1.24	0.1311
BL-8 - BL-9A	-3.57	0.0003	-16.78	< 0.0001	-15.76	< 0.0001
BL-1 - BL-9B	-6.44	< 0.0001	3.32	0.0006	1.58	0.0840
BL-2 - BL-9B	-6.49	< 0.0001	3.14	0.0011	1.65	0.0770
BL-3 - BL-9B	4.55	< 0.0001	0.70	0.2609	0.27	0.4090
BL-4 - BL-9B	0.71	0.2556	-0.56	0.2996	1.33	0.1233
BL-6 - BL-9B	-4.92	< 0.0001	9.79	< 0.0001	8.79	< 0.0001
BL-8 - BL-9B	-9.00	< 0.0001	-8.06	< 0.0001	-8.49	< 0.0001
BL-9A - BL-9B	-5.32	< 0.0001	8.82	< 0.0001	7.37	< 0.0001

Table 4.3: Pairwise comparisons of the Shannon index between laboratories using Dunn's test for each dataset. Data source: [Sinha et al., 2015].

the pairwise Bray-Curtis distances between samples using a combined taxa matrix which consists of the unfiltered taxa matrix, and the taxa filtered ma-

trices of PERFect simultaneous and PERFect permutation each at the p-value threshold of 0.1. The multidimensional scaling ordination plot for the first two principal components which explain 30.5% of the variability in the data is shown in Figure 4.7. Three filtering methods (unfiltered, simultaneous and permutation PERFect) are arranged in columns and samples are colored according to 8 processing institutes. Figure 4.7 shows that while data clusters by laboratory in each dataset, the proximity between clusters decreases when simultaneous or permutation filtering is applied. This observation indicates that filtering improves similarity between samples that contain the same mock communities and alleviates the effects of lab-to-lab variability.

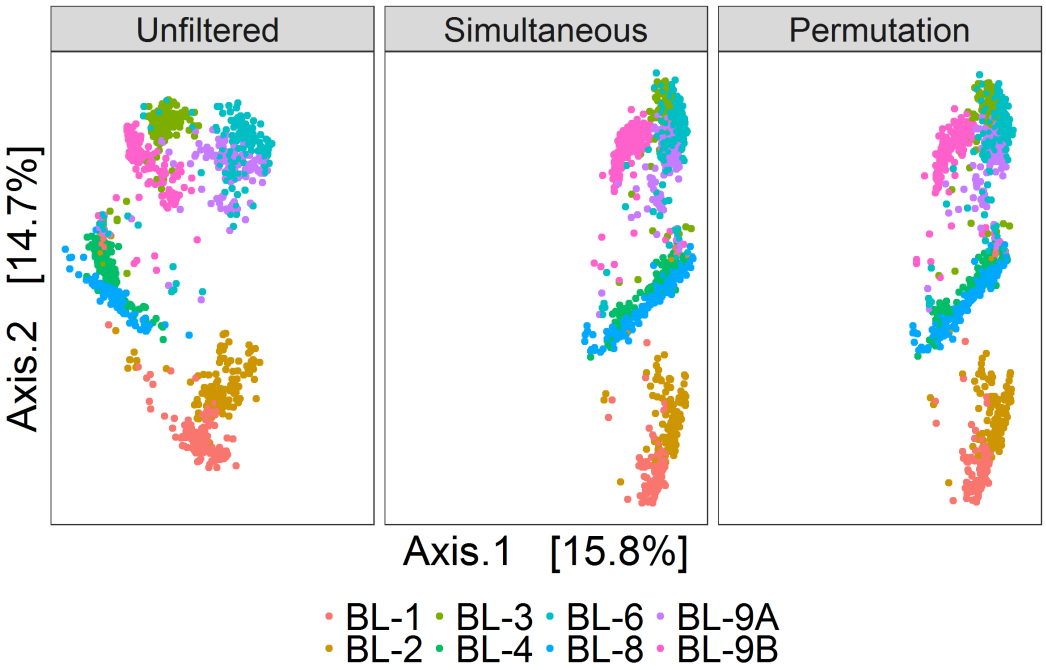


Figure 4.7: Multidimensional scaling plots of the unfiltered, simultaneous and permutation PERFect filtered data colored by bioinformatics processing institutes. Data source: [Sinha et al., 2015].

4.3.2 The reagent and laboratory contamination data

The [Salter et al., 2014] study was designed to determine the effects of DNA extraction kits and other laboratory reagent contamination on sequencing output. These data contain mock samples of a pure Salmonella bongori culture that had been processed at three different institutes: (1) Imperial College London (ICL); (2) University of Birmingham (UB); and (3) Wellcome Trust Sanger Institute (WTSI). Each mock sample underwent five rounds of serial ten-fold dilutions to generate a series of high (dilution = 0) to low (dilution = 5) biomass samples. Figure 4.8 displays the log-counts heat map for 635 observed taxa, generated using 40 Polymerase Chain Reaction (PCR) cycles. The taxa on the horizontal axis are arranged in decreasing order of abundance and the 18 samples on the vertical axis arranged by low to high (0 to 5) degrees of dilution. Figure 4.9 displays the Bray-Curtis distance MDS plots for the 18 samples from the heat map on the left. The first two principal components that explain 81.1% of the variability in the data are shown on the plot. As the dilution number increases, true taxa contain less signal and are observed in lower counts, which makes it difficult to separate the signal from the noise.

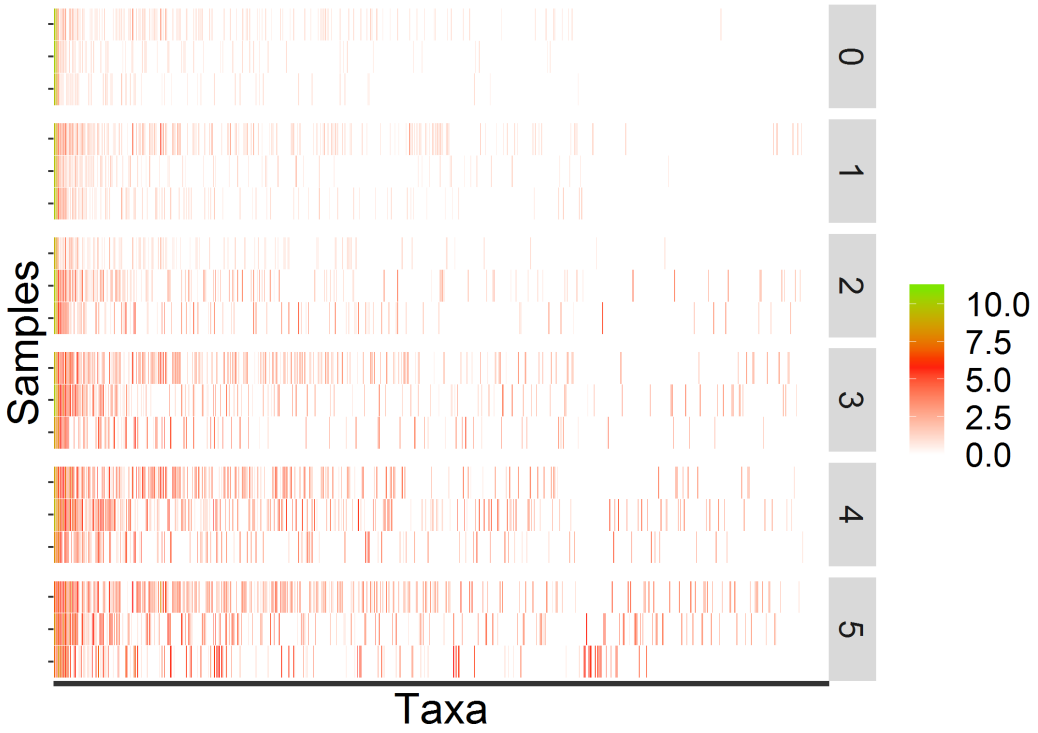


Figure 4.8: The heatmap of the observed taxa on the log-scale, with taxa on the x-axis arranged in decreasing abundance order and samples on the y-axis arranged from low to high (0 to 5) degrees of dilution. Source: [Salter et al., 2014]

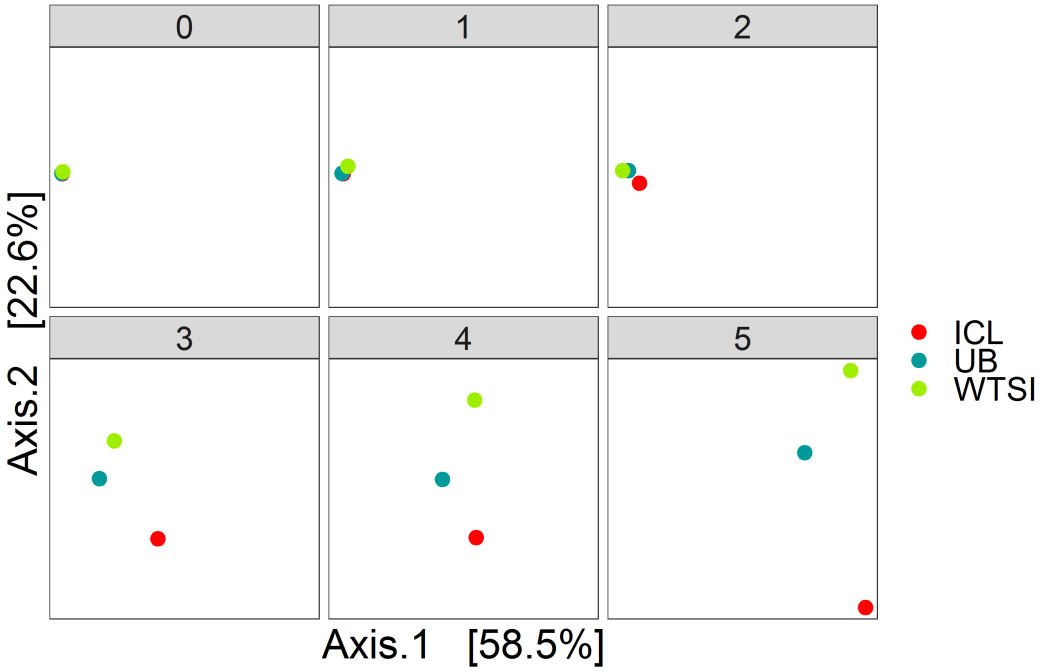


Figure 4.9: The multidimensional scaling plots for each degree of dilution, colored by the processing institutes. Source: [Salter et al., 2014]

Figure 4.10 displays the difference in the alpha diversity of the filtered outputs, corresponding to the p-value threshold 0.1, using simultaneous and permutation filtering among 6 dilution levels and 3 processing institutes. The left panel uses the log of the Chao1 index to measure the richness of species in the unfiltered and filtered data, colored by the dilution levels. The original data have the highest Chao1 indices which increase with dilution levels, indicating that there are more contaminants in higher dilution samples. The PERFect simultaneous method preserves the least number of species, which leads to the least richness and variability due to the dilution levels. Both PERFect filtering methods alleviate the increasing trend with dilution levels (clearer effect in the Simultaneous method), thus we conclude that filtering successfully mitigates the dilution effects measured by Chao1 index. The right panel shows

the Shannon index, color by the dilution number. It is expected that as the dilution levels increase, the proportion of signal taxa decreases whereas that of noise taxa increases, causing the data to become more even, and thus leading to a higher Shannon index. This effect may cause problems comparing alpha diversity for different groups of samples with variable biomass because it will be more difficult to differentiate between signal and noise taxa in low biomass samples. The filtering methods address this issue by removing noise taxa in highly diluted samples (dilutions $3, 4$ and 5), where the simultaneous filtering removes more taxa than the permutation algorithm and has more impact on reducing the alpha diversity.

To compare the beta diversity for filtered outputs, the pairwise between-sample Bray-Curtis distances are calculated using the taxa matrices' combination with a similar set up to the analysis with the MBQC data. The multidimensional scaling ordination plot for the first two principal components that explain 81.3% of the variability in the data is shown in Figure 4.11. The six dilution levels and three filtering methods (none, simultaneous and permutation PERFect) are arranged in columns and rows respectively; samples are colored according to the three processing institutes. Ideally, the samples from all three processing institutes should have the same composition of taxa regardless of the dilution levels. However, contaminants that went into the samples during the DNA extraction and PCR process lead to the differences between the three processing institutes. Figure 4.11 shows that filtering does not dramatically change samples' pairwise distances in ordination plots. This is due to the fact that PERFect, like many other filtering methods, removes taxa with low abundance which do not contribute to the signal, and thus do not dramatically affect samples' pairwise distances. These observations lead to

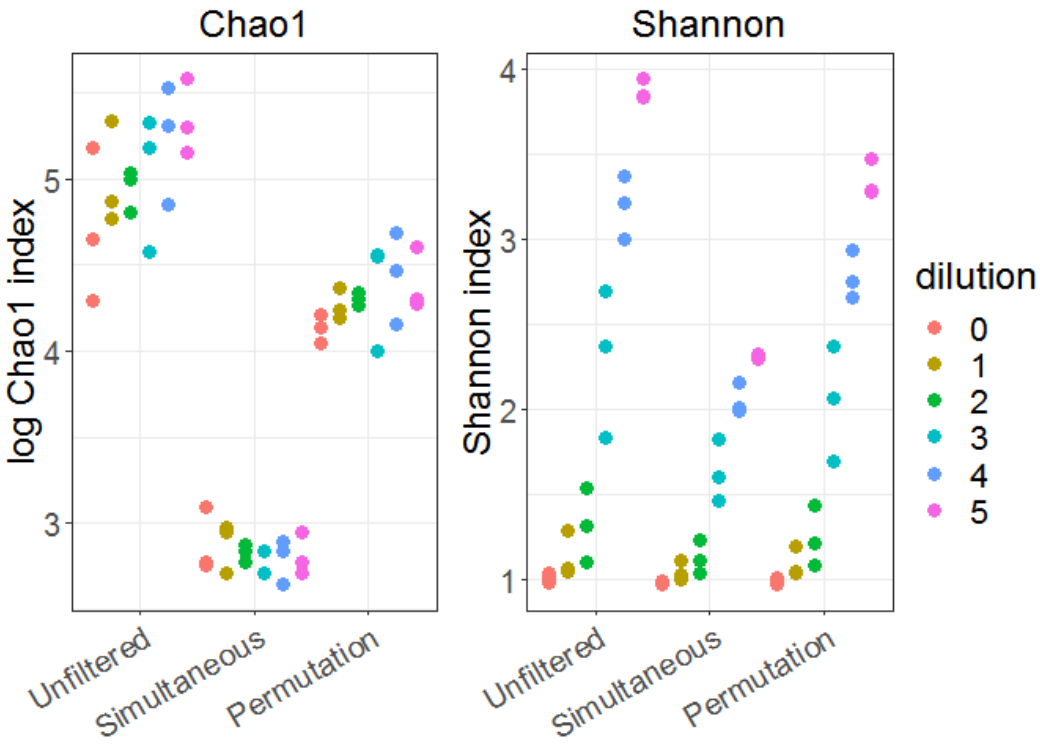


Figure 4.10: Left panel: The logarithm of the Chao1 index for the original data and two filtered data, colored by the dilution levels. Right panel: The Shannon index for the original data and two filtered data, colored by the dilution levels. Data source: [Salter et al., 2014].

the important conclusion that filtering reduces the number of taxa considered in the analysis, and thus reduces dimensionality of the OTU table, without affecting beta diversity.

4.3.3 Computation time

Table 4.4 displays the computation time of the PERFect filtering methods, using the abundance and p-values ordering, on the [Sinha et al., 2015] and [Salter et al., 2014] data. The simultaneous filtering method is the most com-

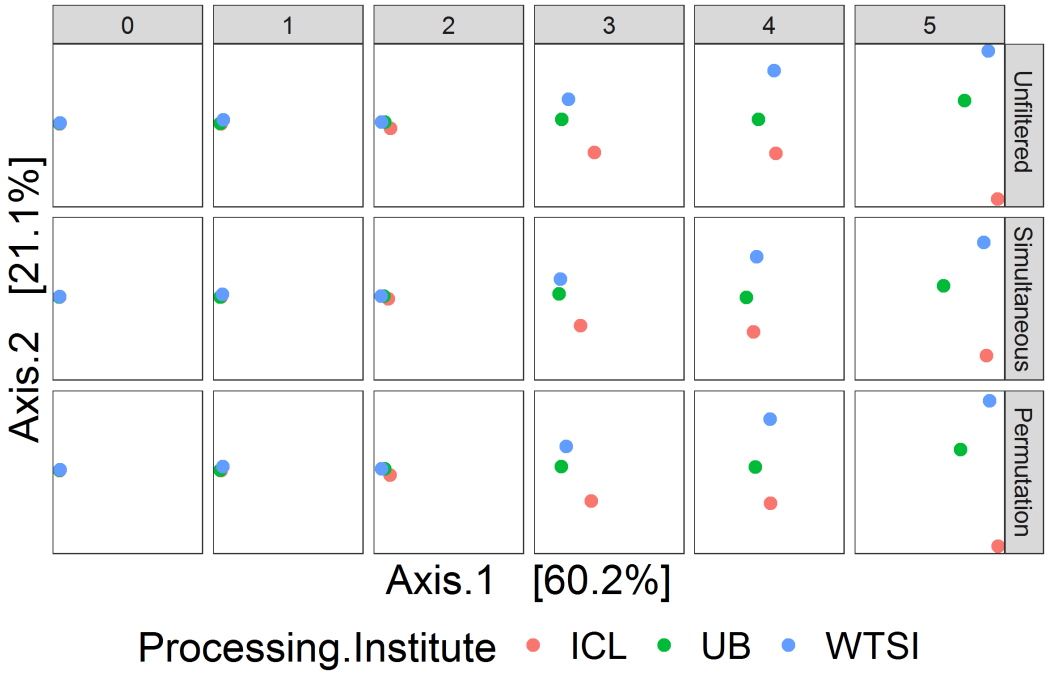


Figure 4.11: Multidimensional scaling plots of the unfiltered and filtered data at different dilution levels, colored by the processing institutes. Data source: [Salter et al., 2014].

putationally efficient because it fits the differences in filtering loss for all taxa using one skew-normal distribution. Although arbitrarily there is no difference in the running time between the two types of ordering, there is a large difference in the running time between the full and fast permutation algorithm. By only calculating the "necessary" taxa's p-values, we have effectively reduced the running time by almost four times. Moreover, it is clear that as the dataset gets larger, the permutation method requires more time to execute the full algorithm, whereas the fast implementation requires much less additional time. Therefore, we set the default of our permutation method to use the fast algorithm (used to conduct the comparisons in section 4.3), but users may choose to use the full algorithm to evaluate the significance of every taxon.

Dataset	Method	Ordering	Computation time
MBQC (1016 × 792)	Simultaneous	Abundance	1.38
	Simultaneous	P-values	1.18
	Permutation (full algorithm)	Abundance	750.39
	Permutation (fast algorithm)	Abundance	212.14
	Permutation (full algorithm)	P-values	808.75
	Permutation (fast algorithm)	P-values	216.60
Salter (42 × 635)	Simultaneous	Abundance	0.50
	Simultaneous	P-values	0.45
	Permutation (full algorithm)	Abundance	462.70
	Permutation (fast algorithm)	Abundance	153.30
	Permutation (full algorithm)	P-values	448.25
	Permutation (fast algorithm)	P-values	153.66

Table 4.4: The running time (in seconds) for different settings of the filtering methods on the [Sinha et al., 2015] and [Salter et al., 2014] data.

Chapter 5

Final Remarks

In this dissertation, we introduce mortality models using reliable measures derived from accelerometry data. A total of 33 predictors of 5-year all-cause mortality, including 20 measures of objective PA, were compared using univariate and multivariate logistic regression. In univariate logistic regression, the total activity count was the best predictor of 5-year mortality (AUC = 0.771), followed by age (AUC = 0.758). Overall, 9 of the top 10 predictors were objective PA measures (AUC from 0.771 to 0.692). Hence, objective accelerometry-derived PA measures have the potential to outperform traditional predictors of 5-year mortality, including age. This highlights the importance of wearable technology for providing reproducible, unbiased, and prognostic biomarkers of health.

In multiple regression, the best 10-fold cross-validated AUC was 0.798 for the model without objective PA variables (9 predictors) and 0.838 for the forward selection model with objective PA variables (13 predictors, including two PA variables which are the ASTP and the surrogate of the standard deviation of the PC6). Here, we use these surrogates for the models since they make

it possible to do predictions for new data (not originally observed in the prediction model). Indeed, if we collect activity for another person who did not participate in the NHANES study and would like to use the final mortality prediction model to predict their 5-year mortality risk, we would have to recalculate the PCs using both the NHANES and the new person data. This is clearly not practical. Thus, we replace the PC scores by the surrogates which can be easily calculate from the new data.

A limitation of the forward selection model is the exclusion of interaction terms from the analysis. Unreported results indicate that most predictive interactions were between age and objective PA predictors. Although some of the interactions were significant, they did not fundamentally change the results. Thus, to preserve simplicity, the focus here is on main effects prediction. Another limitation is our focus on maximizing cross-validated AUC and we did not examine the change in contribution of one variable after accounting for others in the model. In the future, we plan to perform stepwise variable selection and compare its result to the current model. Finally, the association between physical activity and time to death using Cox models in the NHANES population has been investigated in several publications [Fishman et al., 2016a], [Di et al., 2017], [Leroux et al., 2019] and the results are consistent with our findings in this dissertation.

In the future work, we plan to investigate Cox modeling, other binary endpoints, such as 1-, 2-, 3-, 4- year mortality and cause-specific mortality. The unique perspective our method provides is the focus on quantifying the absolute and relative performance of mortality predictors. We hope that this will add clarity to the large, existing literature on mortality predictors and will provide much needed information about the individual, combined, and

relative prediction performance of these predictors. We also hope to illustrate the strong performance of objectively measured physical activity predictors of mortality relative to well-known, established predictors. We further plan to build mortality models using the UK Biobank data (the equivalent of US NHANES data in the UK), validating interactions between objective PA predictors and other covariates and studying the effect of changing/extending the prediction horizon.

For the filtering problem, there is no single statistical method besides the rule of thumb that suggests to keep taxa that are present in at least k samples. Another popular approach is to remove taxa that are observed in fewer than $k\%$ of the samples. The advantage of these methods is that they are simple, intuitive, and easy to communicate with collaborators. However, they do not have an explicit loss function and objective criteria for choosing the tuning parameters m and k. The method **decontam** identifies contaminants by: 1) inversely correlating taxa frequencies with sample DNA concentration; and 2) using the prevalence of sequenced negative controls but the auxiliary data might not always be available. In this dissertation, we introduce the **R** package **PERFect** implementing the PERFect microbiome filtering algorithm, which is a statistically driven method for choosing a threshold value for removing rare taxa [Smirnova et al., 2018a]. A comparison of these methods (two rules of thumb, **decontam** and **PERFect**) is performed by [Smirnova et al., 2018a] and results show that **PERFect** removes rare taxa and contaminants more effectively. We then further improved our method by implementing a fast version of the permutation algorithm, which significantly reduces computational time. In addition, we focus on the improvement of alpha and beta diversity estimation and reduction of lab-to-lab variability between samples that contain similar

microbial species. Results indicate strong potential of the filtering methods in alleviating the differences between samples processed at different institutes and according to different protocols. Moreover, filtering removes rare taxa that have low contribution to the signal, thus reducing dimensionality of the data with minimal information loss. To the best of our knowledge, *this is the first report on the effects of the PERFect filtering approach on downstream analyses of microbiome data*.

A limitation of filtering is that the reduction of type I errors (probability of removing important taxa) will inevitably increase type II errors (probability of keeping unimportant taxa). Indeed, if we want to be cautious in removing rare taxa to ensure that important taxa will still remain in the data, we will not remove many taxa and will likely have a lot of unimportant taxa remained; if we remove taxa aggressively, there is a high chance of filtering important rare taxa. In particular, in researches that aim to study rare taxa, filtering would not be advisable since it will likely remove the rare but important taxa. However, this is a general limitation of any filtering approach that does not consider additional information about negative controls, or feature DNA concentrations in the samples. This issue can be moderated by having a good understanding of the data (where the data are sampled and how they are generated) and using auxiliary data that allows us to filter with confidence. In the future package implementation, we will incorporate this information in our filtering approach, as well as relax the taxa ordering assumption.

In the comparison of data quality between labs, if we have greater confidence in some labs' protocols over others, a possible way to reduce noise in the data before filtering is to consider a weighted average of data from the labs, where labs with higher confidence receive more weights. Specifically, data from

all the labs are combined using the weighted mean of the corresponding taxon count and the final weighted taxa matrix will be used for filtering.

Finally, in this dissertation, we evaluated the effect of filtering on the descriptive microbiome analyses. In our future work, we plan to evaluate the effects of filtering on case-control group comparison methods, such as random forest model, linear discriminant analysis effect size (LEfSe) and disease association (biomarker discovery) problems.

Bibliography

[Aguilera and Aguilera-Morillo, 2013] Aguilera, A. and Aguilera-Morillo, M. (2013). Comparative study of different b-spline approaches for functional data. *Mathematical and Computer Modelling*, 58(7):1568 – 1579.

[Albenberg et al., 2012] Albenberg, L. G., Lewis, J. D., and Wu, G. D. (2012). Food and the gut microbiota in inflammatory bowel diseases. *Current Opinion in Gastroenterology*, 28(4):314–320.

[Augustin et al., 2017] Augustin, N. H., Mattocks, C., Faraway, J. J., Greven, S., and Ness, A. R. (2017). Modelling a response as a function of high-frequency count data: The association between physical activity and fat mass. *Statistical Methods in Medical Research*, 26(5):2210–2226. PMID: 26187735.

[Azzalini, 2005] Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188.

[Bai et al., 2016] Bai, J., Di, C., Xiao, L., Evenson, K. R., LaCroix, A. Z., Crainiceanu, C. M., and Buchner, D. M. (2016). An activity index for raw accelerometry data and its comparison with other activity metrics. *PLOS ONE*, 11(8):1–14.

[Bai et al., 2012] Bai, J., Goldsmith, J., Caffo, B., Glass, T. A., and Crainiceanu, C. M. (2012). Movelets: A dictionary of movement.

[Bai et al., 2013] Bai, J., He, B., Shou, H., Zipunnikov, V., Glass, T. A., and Crainiceanu, C. M. (2013). Normalization and extraction of interpretable metrics from raw accelerometry data. *Biostatistics*, 15(1):102–116.

[Belizario and Napolitano, 2015] Belizario, J. E. and Napolitano, M. (2015). Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches. *Frontiers in Microbiology*, 6:1050.

[Bosq, 2000] Bosq, D. (2000). *Linear Processes in Function Spaces; Theory and Applications*. Springer.

[Brooks et al., 2015] Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., Vaginal Microbiome Consortium (additional members), Strauss, J. F., Jefferson, K. K., and Buck, G. A. (2015). The truth about metagenomics: quantifying and counteracting bias in 16s rrna studies. *BMC Microbiology*, 15(1):66.

[Brown et al., 2013] Brown, D. D., Kays, R., Wikelski, M., Wilson, R., and Klimley, P. (2013). Observing the unwatchable through acceleration logging of animal behavior.

[Cacho et al., 2015] Cacho, A., Smirnova, E., Huzurbazar, S., and Cui, X. (2015). A comparison of base-calling algorithms for illumina sequencing technology. *Briefings in Bioinformatics*, 17(5):786âĂŞ795.

[Callahan et al., 2017] Callahan, B. J., DiGiulio, D. B., Goltsman, D. S. A., Sun, C. L., Costello, E. K., Jeganathan, P., Biggio, J. R., Wong, R. J., Druzin, M. L., Shaw, G. M., Stevenson, D. K., Holmes, S. P., and Relman, D. A. (2017). Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of us women. *Proceedings of the National Academy of Sciences*, 114(37):9966–9971.

[CDC, 2016] CDC (2016). About the national health and nutrition examination survey. http://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

[Chen et al., 2012] Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113.

[Choudhury et al., 2019] Choudhury, P. P., Wilcox, A. N., Brook, M. N., Zhang, Y., Ahearn, T., Orr, N., Coulson, P., Schoemaker, M. J., Jones, M. E., Gail, M. H., and et al. (2019). Comparative validation of breast cancer risk prediction models and projections for future risk stratification. *JNCI: Journal of the National Cancer Institute*.

[Croux and Ruiz-Gazen, 2005] Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206 – 226.

[Curtin et al., 2013] Curtin, L. R., Mohadjer, L. K., Dohrmann, S. M., Kruszon-Moran, D., Mirel, L. B., Carroll, M. D., Hirsch, R., Burt, V. L., and Johnson, C. L. (2013). National health and nutrition examination survey: sample design, 2007-2010.

[Davis et al., 2018] Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, 6(1):226.

[de Boor, 2001] de Boor, C. (2001). *A Practical Guide to Splines*. Springer.

[Di et al., 2017] Di, J., Leroux, A., Urbanek, J., Varadhan, R., Spira, A. P., Schrack, J., and Zipunnikov, V. (2017). Patterns of sedentary and active time accumulation are associated with mortality in us adults: The nhanes study. *BioRxiv*.

[DiGiulio et al., 2015] DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., Sun, C. L., Goltsman, D. S. A., Wong, R. J., Shaw, G., Stevenson, D. K., Holmes, S. P., and Relman, D. A. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences*, 112(35):11060–11065.

[Fan and Lv, 2008] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

[Fettweis et al., 2012] Fettweis, J., Serrano, M., Sheth, N., Mayer, C., Glascock, A., Brooks, J., Jefferson, K., Vaginal Microbiome Consortium, a., and Buck, G. (2012). Species-level classification of the vaginal microbiome. *BMC Genomics*, 13(Supplement 18):1–9.

[Finch, 2013] Finch, C. F. (2013). Applications of functional data analysis: A systematic review.

[Fishman et al., 2016a] Fishman, E. I., Steeves, J. A., Zipunnikov, V., Koster, A., Berrigan, D., Harris, T. A., and Murphy, R. (2016a). Association between objectively measured physical activity and mortality in nhanes. *Medicine and Science in Sports and Exercise*, 48(7):1303âĂŞ1311.

[Fishman et al., 2016b] Fishman, E. I., Steeves, J. A., Zipunnikov, V., Koster, A., Berrigan, D., Harris, T. A., and Murphy, R. (2016b). Association between objectively measured physical activity and mortality in nhanes. *Medicine and Science in Sports and Exercise*, 48(7):1303âĂŞ1311.

[Goldsmith et al., 2018] Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2018). *refund: Regression with Functional Data*. R package version 0.1-17.

[Happ and Greven, 2015] Happ, C. and Greven, S. (2015). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659.

[HMP, 2008] HMP (2008). About the human microbiome. https://hmpdacc.org/hmp/overview/. Accessed: 2019-10-29.

[Huttenhower et al., 2014] Huttenhower, C., Kostic, A., and Xavier, R. (2014). Inflammatory bowel disease as a model for translating the microbiome. *Immunity*, 40(6):843–854.

[JHU, 2019] JHU (2019). Wearable activity trackers a reliable tool for predicting death risk in older adults.

[Johnstone and Titterington, 2009] Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data.

[Karas et al., 2019] Karas, M., Bai, J., Aczkiewicz, M., Harezlak, J., Glynn, N. W., Harris, T., Zipunnikov, V., Crainiceanu, C., and Urbanek, J. K. (2019). Accelerometry data in health research: Challenges and opportunities. *Statistics in Biosciences*, 11(2):210–237.

[Knights et al., 2011] Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature methods*, 8(9):761–763.

[Lahti et al., 2013] Lahti, L., Salonen, A., Kekkonen, R. A., SalojÃďrvi, J., Jalanka-Tuovinen, J., Palva, A., OreÅąiÄĐ, M., and De Vos, W. M. (2013). Associations between the human intestinal microbiota,lactobacillus rhamnosusgg and serum lipids indicated by integrated analysis of high-throughput profiling data. *PeerJ*, 1.

[Leite et al., 2019] Leite, G., Villanueva-Millan, M. J., Celly, S., Sedighi, R., Morales, W., Rezaie, A., Weitsman, S., Mathur, R., Parodi, G., Sanchez, M., and et al. (2019). First large scale study defining the characteristic microbiome signatures of small intestinal bacterial overgrowth (sibo): Detailed analysis from the reimagine study. *Gastroenterology*, 156(6).

[Leroux et al., 2019] Leroux, A., Di, J., Smirnova, E., Mcguffey, E., Cao, Q., Bayat, E., Tabacu, L., Zipunnikov, V., Urbanek, J., and Crainiceanu, C. (2019). Organizing and analyzing the activity data in nhanes. *Statistics in Biosciences*, 11.

[Li and Homer, 2010] Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473âĂŞ483.

[Lumley and Scott, 2015] Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1):1–18.

[Matsen, 2014] Matsen, Frederick A., I. (2014). Phylogenetics and the Human Microbiome. *Systematic Biology*, 64(1):e26–e41.

[Maurice et al., 2013] Maurice, C. F., Haiser, H. J., and Turnbaugh, P. J. (2013). Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell*, 152(1):39 – 50.

[Morin, 2013] Morin, P. (2013). *Open Data Structures: An Introduction*. Athabasca University Press.

[NCHS, 2015] NCHS (2015). Office of analysis and epidemiology, public-use linked mortality file.

[Nguyen et al., 2015] Nguyen, L. D. N., Viscogliosi, E., and Delhaes, L. (2015). The lung mycobiome: an emerging field of the human respiratory microbiome. *Frontiers in Microbiology*, 6:89.

[Park and Allaby, 2017a] Park, C. and Allaby, M. (2017a). alpha diversity.

[Park and Allaby, 2017b] Park, C. and Allaby, M. (2017b). beta diversity.

[Pascale et al., 2019] Pascale, A., Marchesi, N., Govoni, S., Coppola, A., and G aruso, C. (2019). The role of gut microbiota in obesity, diabetes mellitus,

and effect of metformin: new insights into old diseases. *Current Opinion in Pharmacology*, 49:1–5.

[Paulson et al., 2017] Paulson, J. N., Talukder, H., and Bravo, H. C. (2017). Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines. *Nature Methods*.

[Pencina et al., 2010] Pencina, M. J., Dagostino, R. B., and Steyerberg, E. W. (2010). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*, 30(1):11âĂŞ21.

[PERFect, 2019] PERFect (2019). Perfect (development version).

[Petrosino, 2018] Petrosino, J. F. (2018). The microbiome in precision medicine: the way forward.

[Proctor, 2014] Proctor, L. M. (2014). The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host and microbe*, 16(3):276–289.

[Puri et al., 2018] Puri, P., Liangpunsakul, S., Christensen, J. E., Shah, V. H., Kamath, P. S., Gores, G. J., Walker, S., Comerford, M., Katz, B., Borst, A., Yu, Q., Kumar, D. P., Mirshahi, F., Radaeva, S., Chalasani, N. P., Crabb, D. W., and Sanyal, A. J. (2018). The circulating microbiome signature and inferred functional metagenomics in alcoholic hepatitis. *Hepatology*, 67(4):1284–1302.

[Quaak and Kuiper, 2011] Quaak, F. C. and Kuiper, I. (2011). Statistical data analysis of bacterial t-rflp profiles in forensic soil comparisons. *Forensic Science International*, 210(1-3):96–101.

[Ramsay and Silverman, 2006] Ramsay, J. O. and Silverman, B. W. (2006). *Functional Data Analysis*. Springer, second edition.

[Ramsay et al., 2018] Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2018). *fda: Functional Data Analysis*. R package version 2.4.8.

[Rao, 1958] Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17.

[Ravel et al., 2011] Ravel, J., Gajer, P., Abdo, Z., Schneider, G., Koenig, S. K., McCulle, S., Karlebach, S., Gorle, R., Russell, J., Tacket, C., Brotman, R., Davis, C., Ault, K., Peralta, L., and Forney, L. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4680–4687.

[Reese and Dunn, 2018] Reese, A. T. and Dunn, R. R. (2018). Drivers of microbiome biodiversity: A review of general rules, feces, and ignorance. *mBio*, 9(4).

[Salter et al., 2014] Salter, S., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N., Walker, A. W., and et al. (2014). Reagent contamination can critically impact sequence-based microbiome analyses. *BMC Biology*.

[Sanschagrin and Yergeau, 2014] Sanschagrin, S. and Yergeau, E. (2014). Next-generation sequencing of 16s ribosomal rna gene amplicons.

[Shinohara et al., 2011] Shinohara, R. T., Crainiceanu, C. M., Caffo, B. S., GaitÃąn, M. I., and Reich, D. S. (2011). Population-wide principal

component-based quantification of blood-brain-barrier dynamics in multiple sclerosis. *NeuroImage*, 57(4):1430 – 1446.

[Sinha et al., 2015] Sinha, R., Abnet, C. C., White, O., Knight, R., and Huttenhower, C. (2015). The microbiome quality control project: baseline study design and future directions. *Genome Biology*, 16(1):276.

[Smirnova et al., 2018a] Smirnova, E., Huzurbazar, S., and Jafari, F. (2018a). Perfect: Permutation filtering test for microbiome data. *Biostatistics*.

[Smirnova et al., 2018b] Smirnova, E., Ivanescu, A., Bai, J., and Crainiceanu, C. M. (2018b). A practical guide to big data. *Statistics and Probability Letters*, 136:25 – 29. The role of Statistics in the era of big data.

[Smirnova et al., 2019] Smirnova, E., Leroux, A., Cao, Q., Tabacu, L., Zipunnikov, V., Crainiceanu, C., and Urbanek, J. K. (2019). The Predictive Performance of Objective Measures of Physical Activity Derived From Accelerometry Data for 5-Year All-Cause Mortality in Older Adults: National Health and Nutritional Examination Survey 2003-2006. *The Journals of Gerontology: Series A*. glz193.

[Solo et al., 2018] Solo, V., Poline, J., Lindquist, M. A., Simpson, S. L., Bowman, F. D., Chung, M. K., and Cassidy, B. (2018). Connectivity in fmri: Blind spots and breakthroughs. *IEEE Transactions on Medical Imaging*, 37(7):1537–1550.

[SPSS, 2018] SPSS (2018). Spss: 50 years of innovation. https://developer.ibm.com/predictiveanalytics/2018/04/05/spss-50-years-innovation. Accessed: 2019-10-29.

[Tolimieri et al., 1989] Tolimieri, R., An, M., and Lu, C. (1989). *Cooley-Tukey FFT Algorithms*, pages 72–93. Springer New York, New York, NY.

[Troiano et al., 2014] Troiano, R. P., McClain, J. J., Brychta, R. J., and Chen, K. Y. (2014). Evolution of accelerometer methods for physical activity research.

[Turnbaugh et al., 2007] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project.

[Usset et al., 2016] Usset, J., Staicu, A.-M., and Maity, A. (2016). Interaction models for functional regression. *Computational Statistics and Data Analysis*, 94:317 – 329.

[Varma et al., 2017] Varma, V. R., Dey, D., Leroux, A., Di, J., Urbanek, J., Xiao, L., and Zipunnikov, V. (2017). Re-evaluating the effect of age on physical activity over the lifespan. *Preventive Medicine*, 101:102 – 108.

[Varma et al., 2018] Varma, V. R., Dey, D., Leroux, A., Di, J., Urbanek, J., Xiao, L., and Zipunnikov, V. (2018). Total volume of physical activity: Tac, tlac or tac(Îż). *Preventive Medicine*, 106:233–235.

[Xia and Sun, 2017] Xia, Y. and Sun, J. (2017). Hypothesis testing and statistical analysis of microbiome. *Genes and Diseases*, 4(3):138 – 148.

[Xiao et al., 2014] Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2014). Fast covariance estimation for high-dimensional functional data. *Statistics and Computing*, 26(1-2):409–421.

[Zhang et al., 2018] Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., and Yi, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in Microbiology*, 9.

Appendix A

Reference Manual for PERFect software

Package 'PERFect'

November 3, 2019

Type Package

Title Permutation filtration for microbiome data

Version 1.1.0

Date 2019-09-02

Author Ekaterina Smirnova <ekaterina.smirnova@vcuhealth.org>,
Quy Cao <quy.cao@umontana.edu>

Maintainer Quy Cao <quy.cao@umontana.edu>

Description PERFect is a novel permutation filtering approach designed to address two unsolved problems in microbiome data processing: (i) define and quantify loss due to filtering by implementing thresholds, and (ii) introduce and evaluate a permutation test for filtering loss to provide a measure of excessive filtering.

Depends R (>= 3.6.0), sn (>= 1.5.2)

Imports ggplot2 (>= 3.0.0), phyloseq (>= 1.28.0), zoo (>= 1.8.3), psych (>= 1.8.4), stats (>= 3.6.0), Matrix (>= 1.2.14), fitdistrplus (>= 1.0.12), parallel (>= 3.6.0)

License Artistic-2.0

Encoding UTF-8

LazyData false

BugReports https://github.com/cxquy91/PERFect/issues

URL https://github.com/cxquy91/PERFect

Suggests knitr, rmarkdown, kableExtra, ggpubr

biocViews Software, Microbiome, Sequencing, Classification, Metagenomics

VignetteBuilder knitr

RoxygenNote 6.1.1

NeedsCompilation no

git_url https://git.bioconductor.org/packages/PERFect

git_branch master

git_last_commit f515877

git_last_commit_date 2019-10-29

Date/Publication 2019-11-02

R topics documented:

`DiffFiltLoss` *Difference in filtering loss*

Description

This function calculates differences in filtering loss due to removing a set of J taxa sequentially.

Usage

```
DiffFiltLoss(X, Order_Ind, Plot = TRUE, Taxa_Names = NULL)
```

Arguments

`X`	OTU table, where taxa are columns and samples are rows of the table. It should be a in dataframe format with columns corresponding to taxa names.
`Order_Ind`	Numeric column order corresponding to taxa importance arrangement.
`Plot`	A binary TRUE/FALSE value. If TRUE, the function returns plot of sequential differences in filtering loss.
`Taxa_Names`	Optional taxa labels corresponding to the columns ordering given by Order_Ind.

Details

This function calculates and plots (if Plot = TRUE) differences in filtering loss sequentially for removing the first j taxa as DFL(j+1) = FL(J_j+1) - FL(J_j) for taxa j=1, ..., p.

Value

`DFL`	Differences in filtering loss values.
`p_FL`	Plot of the differences in filtering loss.

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. "PERFect: permutation filtration of microbiome data', to be submitted.

See Also

FiltLoss

Examples

```
data(mock2)

# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

#arrange counts in order of increasing number of samples taxa are present in
NP <- NP_Order(Counts)

#obtain numeric column order corresponding to taxa importance arrangment
Order_Ind <- match(NP, names(Prop))
DFL <- DiffFiltLoss(X=Prop, Order_Ind = Order_Ind, Plot = TRUE, Taxa_Names = NP)

#Differences in filtering loss values
DFL$DFL

#Plot of the differences in filtering loss
DFL$p_FL
```

FiltLoss *Filtering Loss*

Description

Sequential filtering loss calculation for removing a set of J_j taxa for J= 1, ..., p.

Usage

```
FiltLoss(X, Order = "NP", Order.user = NULL, type = "Cumu", Plot = TRUE)
```

Arguments

X	OTU table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.
Order	Taxa ordering. The default ordering is the number of occurrences (NP) of the taxa in all samples. Other types of order are number of connected taxa and weighted number of connected taxa, denoted as "NC", "NCw" respectively. More details about taxa ordering are described in Smirnova et al. User can also specify their preference order with Order.user.

`Order.user`	User's taxa ordering. This argument takes a character vector of ordered taxa names.
`type`	Type of filtering loss calculation.
	`"Ind"` Individual taxon's filtering loss FL_u(j)
	`"Cumu"` Cumulative filtering loss FL(J) due to removing a set of taxa J
`Plot`	Binary TRUE/FALSE value. If TRUE, the function returns plot of sequential differences in filtering loss.

Details

The individual filtering loss due to removing one taxon j is defined as:

FL_u(j)= 1- (||X^T_-j X_-j||_F^2/||X^TX||_F^2),

where X_-j is the matrix X without column corresponding to jth taxon and ||Z||_F is the Frobenious norm of a matrix Z.

The cumulative filtering loss due to removing a set of taxa is defined as:

FL(J)= 1- (||X^T_-J X_-J||_F^2\||X^TX||_F^2),

where X_-J is the n x (p-|J|) dimensional matrix obtained by removing the columns indexed by the set J from the data matrix X.

The cumulative filtering loss is calculated sequentially for each set of taxa J_j, j=1, ..., p.

Value

`FL`	Filtering loss values.
`p_FL`	Plot of filtering loss values.

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. "PERFect: permutation filtration of microbiome data", to be submitted.

See Also

`DiffFiltLoss`

Examples

```
data(mock2)

# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

#Calculate cumulative filtering loss
FL <- FiltLoss(X=Prop, Order = "NP", type = "Cumu", Plot = TRUE)
```

```
#Differences in filtering loss values
FL$FL

#Plot of the differences in filtering loss
FL$p_FL
```

FL_J *Filtering Loss for a set of filtered taxa J*

Description

This function calculates filtering loss due to removing a group of J taxa.

Usage

```
FL_J(X, J)
```

Arguments

X	OTU table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.
J	A vector of J taxa to be removed. It must be subset of column names of X.

Value

FL Filtering loss value.

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutation filtration of microbiome data", to be submitted.

See Also

FiltLoss

Examples

```
data(mock2)

# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

#arrange counts in order of increasing number of samples taxa are present in
NP <- NP_Order(Counts)
Counts <- Counts[,NP]
```

```
# Extract the taxa names to be removed
J <- colnames(Counts)[1:30]

#Calculate filtering loss due to removing these taxa
FL_J(Counts,J)
```

---

mock2 *Bias experiment data*

Description

These publicly available data (Brooks et al., 2015) were generated as a part of a study designed to evaluate the bias at each step of the VCU sequencing protocol, namely, DNA extraction, PCR amplification, sequencing and taxonomic classification. Mock community samples were created out of 7 vaginally relevant bacteria by mixing prescribed quantities of cells, with quantities varying across samples according to an experimental design described in Brooks et al, 2015. As opposed to the positive controls data, bacteria appear in different proportions across samples. The number of taxa identified by the sequencing and bioinformatics pipeline was 46.

Usage

```
data(mock2)
```

Format

This file contains a count OTU table and a proportion OTU table, each with 240 samples and 46 taxa. A list of true taxa is also given.

NCw_Order *Taxa importance ordering by the weighted number of connected taxa*

Description

Taxa importance ordering by the weighted number of connected taxa

Usage

```
NCw_Order(Counts)
```

Arguments

Counts	OTU COUNTS table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.

Value

NCW Taxa names in increasing order of the weighted number of connected taxa.

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutation filtration of microbiome data", to be submitted.

Examples

```
data(mock2)
# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

#arrange counts in order of increasing number of samples taxa are present in
NCw <- NCw_Order(Counts)
```

NC_Order *Taxa importance ordering by the number of connected taxa*

Description

Taxa importance ordering by the number of connected taxa

Usage

```
NC_Order(Counts)
```

Arguments

Counts	OTU COUNTS table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.

Value

NC Taxa names in increasing order of the number of connected taxa.

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutation filtration of microbiome data", to be submitted.

Examples

```
data(mock2)
# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

#arrange counts in order of increasing number of samples taxa are present in
NC <- NC_Order(Counts)
```

NP_Order *Taxa importance ordering by the number of occurrences of the taxa in the n samples*

Description

Taxa importance ordering by the number of occurrences of the taxa in the n samples

Usage

```
NP_Order(Counts)
```

Arguments

Counts	OTU COUNTS table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.

Value

NP Taxa names in increasing order of the number of samples taxa are present in.

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutation filtration of microbiome data", to be submitted.

Examples

```
data(mock2)
# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

#arrange counts in order of increasing number of samples taxa are present in
NP <- NP_Order(Counts)
```

PERFect_perm *Permutation PERFect filtering for microbiome data*

Description

Permutation filtering of the provided OTU table X at a test level alpha. Each set of j taxa significance is evaluated by fitting the Skew-Normal, Normal, t or Cauchy distribution to the sampling distribution obtained by permuted taxa labels.

Usage

```
PERFect_perm(X, infocol = NULL, Order = "NP", Order.user = NULL, normalize = "counts",
    algorithm = "fast", center = FALSE, quant = c(0.1, 0.25, 0.5),
   distr = "sn", alpha = 0.1, rollmean = TRUE, direction = "left", pvals_sim = NULL,
    k = 10000, nbins = 30, hist = TRUE, col = "red", fill = "green",
    hist_fill = 0.2, linecol = "blue")
```

Arguments

X	OTU table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.
infocol	Index vector of the metadata. We assume user only gives a taxa table, but if the metadata of the samples are included in the columns of the input, this option needs to be specified.
Order	Taxa ordering. The default ordering is the number of occurrences (NP) of the taxa in all samples. Other types of order are p-value ordering, number of connected taxa and weighted number of connected taxa, denoted as "pvals", "NC", "NCw" respectively. More details about taxa ordering are described in Smirnova et al. User can also specify their preference order with Order.user.
Order.user	User's taxa ordering. This argument takes a character vector of ordered taxa names.
normalize	Normalizing taxa count. The default option does not normalize taxa count, but user can convert the OTU table into a proportion table using the option "prop" or convert it into a presence/absence table using "pres".
algorithm	Algorithm speed. The default is speed is "fast", which allows the program to efficiently search for significant taxa without computing all the p-values. User must use the default option "hist = FALSE" for the fast algorithm. The alternative setting is "full", which computes all the taxa's p-values.
center	Centering OTU table. The default option does not center the OTU table.
quant	Quantile values used to fit the distribution to log DFL values. The number of quantile values corresponds to the number of parameters in the distribution the data is fitted to. Assuming that at least 50% of taxa are not informative, we suggest fitting the log Skew-Normal distribution by matching the 10%, 25% and 50% percentiles of the log-transformed samples to the Skew-Normal distribution.
distr	The type of distribution to fit log DFL values to. While we suggest using Skew-Normal distribution, and set as the default distribution, other choices are available.

	"sn" Skew-Normal distribution with 3 parameters: location xi, scale omega^2 and shape alpha
	"norm" Normal distribution with 2 parameters: mean and standard deviation sd
alpha	Test level alpha, set to 0.1 by default.
rollmean	Binary TRUE/FALSE value. If TRUE, rolling average (moving mean) of p-values will be calculated, with the lag window set to 3 by default.
direction	Character specifying whether the index of the result should be left- or right-aligned or centered compared to the rolling window of observations, set to "left" by default.
pvals_sim	Object resulting from simultaneous PERFect with taxa abundance ordering, allowing user to perform Simultaneous PERFect with p-values ordering. Be aware that the choice of distribution for both methods must be the same.
k	The number of permutations, set to 10000 by default.
nbins	Number of bins used to visualize the histogram of log DFL values, set to 30 by default.
hist	Binary TRUE/FALSE value. If TRUE, the function builds histograms for each taxon.
col	Graphical parameter for color of histogram bars border, set to "red" by default.
fill	Graphical parameter for color of histogram fill, set to "green" by default.
hist_fill	Graphical parameter for intensity of histogram fill, set to 0.2 by default.
linecol	Graphical parameter for the color of the fitted distribution density, set to "blue" by default.

Details

Filtering is the process of identifying and removing a subset of taxa according to a particular criterion. As opposed to the the simultaneous filtering approach, we do not assume that all distributions for each set of taxa are identical and equal to the distribution of simultaneous filtering. Function PERFect_perm() filters the provided OTU table X and outputs a filtered table that contains signal taxa. PERFect_perm() calculates differences in filtering loss DFL for each taxon according to the given taxa order. By default, the function fits Skew-Normal distribution to the log-differences in filtering loss but Normal, t, or Cauchy distributions can be also used.

Value

If "algorithm = full" is chosen, a list is returned containing:

filtX	Filtered OTU table.
info	The metadata information.
pvals	P-values of the test.
DFL	Differences in filtering loss values.
fit	Fitted values and further goodness of fit details passed from the fitdistr() function.
hist	Histogram of log differences in filtering loss.
est	Estimated distribution parameters.
dfl_distr	Plot of differences in filtering loss values.

If "algorithm = fast" is chosen, fit, hist, est, dfl_distr will not be returned.

Author(s)

Ekaterina Smirnova

References

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. Scandinavian Journal of Statistics, 32(2), 159-188.

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutationfiltration of microbiome data", to be submitted.

See Also

PERFect_sim

Examples

```
data(mock2)

# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

# Perform simultaenous filtering of the data
res_sim <- PERFect_sim(X=Counts)

#order according to p-values
pvals_sim <- pvals_Order(Counts, res_sim)

## Not run:
# obtain permutation PERFEct results using NP taxa ordering
res_perm <- PERFect_perm(X = Prop, Order.user = pvals_sim, algorithm = "fast")

# permutation perfect colored by FLu values
pvals_Plots(PERFect = res_perm, X = Counts, quantiles = c(0.25, 0.5, 0.8, 0.9), alpha=0.05)

## End(Not run)
```

PERFect_perm_reorder *Permutation PERFect filtering for microbiome data*

Description

This function filters the provided OTU table X at a test level alpha given a fitted object perfect_perm obtained by running PERFect_perm() function. PERFect_perm_reorder() reavaluates taxa significance p-values for a different taxa ordering.

Usage

```
PERFect_perm_reorder(X, Order = "NP", Order.user = NULL, res_perm, normalize = "counts",
   center = FALSE, alpha = 0.1, distr = "sn", rollmean = TRUE, direction = "left",
    pvals_sim = NULL)
```

Arguments

X	OTU table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.
Order	Taxa ordering. The default ordering is the number of occurrences (NP) of the taxa in all samples. Other types of order are p-value ordering, number of connected taxa and weighted number of connected taxa, denoted as "pvals", "NC", "NCw" respectively. More details about taxa ordering are described in Smirnova et al. User can also specify their preference order with Order.user.
Order.user	User's taxa ordering. This argument takes a character vector of ordered taxa names.
res_perm	Output of PERFect_perm() function.
normalize	Normalizing taxa count. The default option does not normalize taxa count, but user can convert the OTU table into a proportion table using the option "prop" or convert it into a presence/absence table using "pres".
center	Centering OTU table. The default option does not center the OTU table.
alpha	Test level alpha, set to 0.1 by default.
distr	The type of distribution used in PERFect_perm() function to obtain res_perm object.
	"sn" Skew-Normal distribution with 3 parameters: location xi, scale omega^2 and shape alpha
	"norm" Normal distribution with 2 parameters: mean and standard deviation sd
	"t" Student t-distribution with 2 parameters: n degrees of freedom and noncentrality ncp
	"cauchy" Cauchy distribution with 2 parameters: location and scale
rollmean	Binary TRUE/FALSE value. If TRUE, rolling average (moving mean) of p-values will be calculated, with the lag window set to 3 by default.
direction	Character specifying whether the index of the result should be left- or right-aligned or centered compared to the rolling window of observations, set to "left" by default.
pvals_sim	Object resulting from simultaneous PERFect with taxa abundance ordering, allowing user to perform Simultaneous PERFect with p-values ordering. Be aware that the choice of distribution for both methods must be the same.

Details

This function is designed to save computational time needed to obtain and fit the sampling distribution for each taxon if taxa ordering different from the one used in PERFect_perm() is used. Note, the distribution and OTU table X should match the distribution used in PERFect_perm().

Value

res_perm	The perfect_perm object updated according to the alternative taxa ordering. All elements in this list are same as in perfect_perm object given by PERFect() function.

Author(s)

Ekaterina Smirnova

References

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. Scandinavian Journal of Statistics, 32(2), 159-188.

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutationfiltration of microbiome data", to be submitted.

See Also

PERFect_perm

Examples

```
data(mock2)

# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

## Not run:
# obtain permutation PERFEct results using NP taxa ordering
system.time(res_perm <- PERFect_perm(X=Prop, k = 1000, algorithm = "fast"))

# run PERFEct_sim() function and obtain p-values ordering
res_sim <- PERFect_sim(X=Prop)

# order according to p-values
pvals_sim <- pvals_Order(Counts, res_sim)

# update perfect_perm object according to p-values ordering
res_reorder <- PERFect_perm_reorder(X=Prop,  Order.user = pvals_sim,  res_perm = res_perm)

# permutation perfect colored by FLu values
pvals_Plots(PERFect = res_perm, X = Counts, quantiles = c(0.25, 0.5, 0.8, 0.9), alpha=0.05)

## End(Not run)
```

PERFect_sim *Simulation PERFect filtering for microbiome data*

Description

Simultaneous filtering of the provided OTU table X at a test level alpha. One distribution is fit to taxa simultaneously.

Usage

```
PERFect_sim(X,infocol = NULL, Order = "NP", Order.user = NULL, normalize = "counts",
          center = FALSE, quant = c(0.1, 0.25, 0.5), distr = "sn",
          alpha = 0.1, rollmean = TRUE, direction = "left", pvals_sim = NULL,
          nbins = 30, col = "red", fill = "green", hist_fill = 0.2,
          linecol = "blue")
```

Arguments

X	OTU table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names. It could contains columns of metadata.
infocol	Index vector of the metadata. We assume user only gives a taxa table, but if the metadata of the samples are included in the columns of the input, this option needs to be specified.
Order	Taxa ordering. The default ordering is the number of occurrences (NP) of the taxa in all samples. Other types of order are p-value ordering, number of connected taxa and weighted number of connected taxa, denoted as "pvals", "NC", "NCw" respectively. More details about taxa ordering are described in Smirnova et al. User can also specify their preference order with Order.user.
Order.user	User's taxa ordering. This argument takes a character vector of ordered taxa names.
normalize	Normalizing taxa count. The default option does not normalize taxa count, but user can convert the OTU table into a proportion table using the option "prop" or convert it into a presence/absence table using "pres".
center	Centering OTU table. The default option does not center the OTU table.
quant	Quantile values used to fit the distribution to log DFL values. The number of quantile values corresponds to the number of parameters in the distribution the data is fitted to. Assuming that at least 50% of taxa are not informative, we suggest fitting the log Skew-Normal distribution by matching the 10%, 25% and 50% percentiles of the log-transformed samples to the Skew-Normal distribution.
distr	The type of distribution to fit log DFL values to. While we suggest using Skew-Normal distribution, and set as the default distribution, other choices are available. "sn" Skew-Normal distribution with 3 parameters: location xi, scale omega^2 and shape alpha "norm" Normal distribution with 2 parameters: mean and standard deviation sd "t" Student t-distribution with 2 parameters: n degrees of freedom and noncentrality ncp "cauchy" Cauchy distribution with 2 parameters: location and scale
alpha	Test level alpha, set to 0.1 by default.
rollmean	Binary TRUE/FALSE value. If TRUE, rolling average (moving mean) of p-values will be calculated, with the lag window set to 3 by default.
direction	Character specifying whether the index of the result should be left- or right-aligned or centered compared to the rolling window of observations, set to "left" by default.
pvals_sim	Object resulting from simultaneous PERFect with taxa abundance ordering, allowing user to perform Simultaneous PERFect with p-values ordering. Be aware that the choice of distribution for both methods must be the same.
nbins	Number of bins used to visualize the histogram of log DFL values, set to 30 by default.
col	Graphical parameter for color of histogram bars border, set to "red" by default.
fill	Graphical parameter for color of histogram fill, set to "green" by default.
hist_fill	Graphical parameter for intensity of histogram fill, set to 0.2 by default.
linecol	Graphical parameter for the color of the fitted distribution density, set to "blue" by default.

Details

Filtering is the process of identifying and removing a subset of taxa according to a particular criterion. Function `PERFect_sim()` filters the provided OTU table X and outputs a filtered table that contains signal taxa. `PERFect_sim()` calculates differences in filtering loss DFL for each taxon according to the given taxa order. By default, the function fits Skew-Normal distribution to the log-differences in filtering loss but Normal, t, or Cauchy distributions can be also used. This is implementation of Algorithm 1 described in Smirnova et al.

Value

A list is returned containing:

`filtX`	Filtered OTU table.
`info`	The metadata information.
`pvals`	P-values of the test.
`DFL`	Differences in filtering loss values.
`fit`	Fitted values and further goodness of fit details passed from the `fitdistr()` function.
`hist`	Histogram of log differences in filtering loss.
`est`	Estimated distribution parameters.
`pDFL`	Plot of differences in filtering loss values.

Author(s)

Ekaterina Smirnova

References

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. Scandinavian Journal of Statistics, 32(2), 159-188.

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutationfiltration of microbiome data", to be submitted.

See Also

`PERFect_perm`

Examples

```
data(mock2)
# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts
dim(Counts) # 240x46

# Perform simultaenous filtering of the data
res_sim <- PERFect_sim(X=Counts)
dim(res_sim$filtX)      # 240x10, removing 36 taxa
colnames(res_sim$filtX) # signal taxa
```

```
#permutation perfect colored by FLu values
pvals_Plots(PERFect = res_sim, X = Counts, quantiles = c(0.25, 0.5, 0.8, 0.9), alpha=0.05)
```

`pvals_Order` *Taxa importance ordering by PERFect p-values*

Description

This function orders taxa by increasing significance of simultaneous PERFect p-values.

Usage

```
pvals_Order(Counts, res_sim)
```

Arguments

`Counts`	OTU COUNTS table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.
`res_sim`	Output of `PERFect_sim()` function.

Value

Order_pvals Taxa names in increasing order of p-values significance.

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutation filtration of microbiome data", to be submitted.

See Also

`PERFect_sim`, `PERFect_perm`

Examples

```
data(mock2)

# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts

# Perform simultaenous filtering of the data
res_sim <- PERFect_sim(X=Counts)

#order according to p-values
pvals_sim <- pvals_Order(Counts, res_sim)
```

pvals_Plots *Plots of PERFect p-values*

Description

Graphical representation of p-values obtained by running PERFect_sim() or PERFect_perm() for jth taxon colored by quantile values of individual filtering loss.

Usage

```
pvals_Plots(PERFect, X, quantiles = c(0.25, 0.5, 0.8, 0.9), alpha = 0.1)
```

Arguments

PERFect	Output of PERFect_sim() or PERFect_perm() function.
X	OTU table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.
quantiles	Quantile values for coloring, these are set to 25%, 50%, 80% and 90% percentiles of the individual filtering loss values.
alpha	Alpha level of the test, set to 0.1 by default.

Value

A list is returned containing:

df	Dataframe of taxa names, p-values, Flu values and quantiles.
p_vals	Plot of p-values.

Author(s)

Ekaterina Smirnova

See Also

PERFect_sim, PERFect_perm

Examples

```
data(mock2)
# Proportion data matrix
Prop <- mock2$Prop

# Counts data matrix
Counts <- mock2$Counts
dim(Counts) # 240x46

# Perform simultaenous filtering of the data
res_sim <- PERFect_sim(X=Counts)
dim(res_sim$filtX)      # 240x10, removing 36 taxa
colnames(res_sim$filtX) # signal taxa

#permutation perfect colored by FLu values
pvals_Plots(PERFect = res_sim, X = Counts, quantiles = c(0.25, 0.5, 0.8, 0.9), alpha=0.05)
```

TraditR1 *Traditional Filtering Rule 1*

Description

This rule suggests to remove taxa that are mostly absent in all samples.

Usage

```
TraditR1(X, thresh =5)
```

Arguments

X	OTU COUNTS table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.
thresh	Numerical value, set to 5 by default. Throughout all samples, taxa that are present for less than this threshold with be removed.

Value

filtX	Filtered OTU table

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. "PERFect: permutation filtration of microbiome data.

Examples

```
data(mock2)

# Counts data matrix
Counts <- mock2$Counts

# Filtering
Filtered_X <- TraditR1(Counts)
```

TraditR2 *Traditional Filtering Rule 2*

Description

This rule is adopted from Milici et al. (2016) that removes taxa with low abundance level. Specifically, it keeps taxa with abundance level higher than 0.001%. Then it further selects taxa that satisfy at least one of the following conditions: Present in at least one sample at a relative abundance higher than 1% of the reads of that sample, present in at least 2% of samples at a relative abundance higher than 0.1% for a given sample, present in at least 5% of samples at any abundance level.

Usage

```
TraditR2(X, Ab_min = 0.001)
```

Arguments

X	OTU COUNTS table, where taxa are columns and samples are rows of the table. It should be a in data frame format with columns corresponding to taxa names.
Ab_min	Numerical value, set to 0.001 by default. Throughout all samples, taxa with abundance less than this threshold with be removed.

Value

filtX	Filtered OTU table

Author(s)

Ekaterina Smirnova

References

Smirnova, E., Huzurbazar, H., Jafari, F. “PERFect: permutation filtration of microbiome data.

Examples

```
data(mock2)

# Counts data matrix
Counts <- mock2$Counts

# Filtering
Filtered_X <- TraditR2(Counts)
```

Index