Equivalence testing for identity authentication using pulse waves from photoplethysmograph

by

Mengjiao Wu

B.S., Ningbo University, 2012

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

Photoplethysmograph sensors use a light-based technology to sense the rate of blood flow as controlled by the heart's pumping action. This allows for a graphical display of a patient's pulse wave form and the description of its key features. A person's pulse wave has been proposed as a tool in a wide variety of applications. For example, it could be used to diagnose the cause of coldness felt in the extremities or to measure stress levels while performing certain tasks. It could also be applied to quantify the risk of heart disease in the general population. In the present work, we explore its use for identity authentication.

First, we visualize the pulse waves from individual patients using functional boxplots which assess the overall behavior and identify unusual observations. Functional boxplots are also shown to be helpful in preprocessing the data by shifting individual pulse waves to a proper starting point. We then employ functional analysis of variance (FANOVA) and permutation tests to demonstrate that the identities of a group of subjects could be differentiated and compared by their pulse wave forms. One of the primary tasks of the project is to confirm the identity of a person, i.e., we must decide if a given person is whom they claim to be. We used an equivalence test to determine whether the pulse wave of the person under verification and the actual person were close enough to be considered equivalent. A nonparametric bootstrap functional equivalence test was applied to evaluate equivalence by constructing point-wise confidence intervals for the metric of identity assurance. We also proposed new testing procedures, including the way of building the equivalence hypothesis and test statistics, determination of evaluation range and equivalence bands, to authenticate the identity.

Equivalence testing for identity authentication using pulse waves from photoplethysmograph

by

Mengjiao Wu

B.S., Ningbo University, 2012

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Co-Major Professor
Suzanne Dubnicka

Approved by:

Co-Major Professor
Christopher Vahl

Copyright

Abstract

Photoplethysmograph sensors use a light-based technology to sense the rate of blood flow as controlled by the heart's pumping action. This allows for a graphical display of a patient's pulse wave form and the description of its key features. A person's pulse wave has been proposed as a tool in a wide variety of applications. For example, it could be used to diagnose the cause of coldness felt in the extremities or to measure stress levels while performing certain tasks. It could also be applied to quantify the risk of heart disease in the general population. In the present work, we explore its use for identity authentication.

First, we visualize the pulse waves from individual patients using functional boxplots which assess the overall behavior and identify unusual observations. Functional boxplots are also shown to be helpful in preprocessing the data by shifting individual pulse waves to a proper starting point. We then employ functional analysis of variance (FANOVA) and permutation tests to demonstrate that the identities of a group of subjects could be differentiated and compared by their pulse wave forms. One of the primary tasks of the project is to confirm the identity of a person, i.e., we must decide if a given person is whom they claim to be. We used an equivalence test to determine whether the pulse wave of the person under verification and the actual person were close enough to be considered equivalent. A nonparametric bootstrap functional equivalence test was applied to evaluate equivalence by constructing point-wise confidence intervals for the metric of identity assurance. We also proposed new testing procedures, including the way of building the equivalence hypothesis and test statistics, determination of evaluation range and equivalence bands, to authenticate the identity.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Firstly, I would like to express my sincere gratitude to my co-major advisors Dr. Suzanne Dubnicka and Dr. Christopher Vahl for their continuous support of my Ph.D study and related research, for their patience and support in overcoming numerous obstacles I have been facing through my research. Dr. Dubnicka consistently allowed this paper to be my own work, but steered me in the right direction whenever she thought I needed it. The door to Dr. Vahl's office was always open whenever I ran into a trouble spot or had a question about my research or writing.

Besides my advisors, I would like to thank the rest of my committee: Dr. Steven Warren and Dr. Perla Reyes, for their insightful comments and encouragement, for their patience, motivation, and immense knowledge.

I would also like to thank my fellow doctoral students, Kessinee Chitakasempornkul and Yan Wei, for their support, encouragement and of course friendship. In addition I would like to express my gratitude to the staff of Bonnie Messmer and Jo Blackburn for their kindness and long time favors.

I would like to thank my friends for accepting nothing less than excellence from me. Last but not the least, I would like to thank my family: my mother and to my husband and son for providing me unfailing support and continuous encouragement throughout my long Ph.D journey and through the precess of researching and writing this thesis, and my life in general. This accomplishment would not have been possible without them. Thank you.

Chapter 1

Introduction

When we use online or offline services, like register to vote, view our driving records or tax details, apply for an apprenticeship or manage a student loan, we want to be confident that someone else cannot sign in pretending to be us, see our sensitive personal records or use our identity to make fraudulent claims. We want to be confident that our data and services are secure and that our privacy is protected. Here comes the issue of identity authentication. In contrast to traditional authentication, which would prompt a user for the same credentials during every login attempt, continuous biometric identification such as pulse wave, determines the risk associated with the action the user is about to take, based on context gathered over time continuously, then combines that with what is known about the user to make dynamic decisions on how best to authenticate them. In the present work, we explore the use of photoplethysmography (PPG) for identity authentication.

In Chapter 1, we define the user authentication and functional data and then introduce a motivating example of continuous biometric authentication by PPG collected from the pulse oximeter. We also review the existing continuous authentication techniques in Section 1.6.1. However, there are no studies on the use of PPG signals in continuous authentication systems or how to standardize the PPG amplitude for comparing one waveform to another.

To use the pulse waves from PPG for authentication, we need show that the variability among curves between persons is large enough, relative to the variability among curves within people. We visualize the pulse waves using functional boxplot in Section 2.1 and employ the functional analysis of variance (FANOVA) in Section 2.2 and the permutation test in Section 2.3 to differentiate the subjects when the pulse waves are generated from different people.

However, the primary goal of identity authentication could not be achieved by either the FANOVA or the permutation test, or traditional hypothesis testing. We propose a new testing procedure in Chapter 3 to evaluate equivalence for functional data and name it nonparametric bootstrap functional equivalence test (NBFET). In Chapter 4, we simulate the pulse wave-like curves from Brownian motion to determine the appropriate equivalence limits and validate the performance of the proposed testing procedure by computing the power and the type I error rate for various scenarios. Chapter 5 shows the results from real data applications and a summary of our contributions in this work.

1.1 User Authentication

User authentication is the process of verifying whether the identity of a user is genuine before granting him access to resources or services in a secure environment. Traditional authentication methods are based on the user's knowledge, such as personal identification number (PIN) and passwords, or something belonging to someone, such as a smart card or cardkey. It is a well-established fact that traditional passwords are unsafe and may be forgotten, and identification cards can be lost or misplaced. One of the solutions proposed to prevent such occurrences is applying continuous authentication, a new generation of security systems that require the user to re-authenticate themselves in a repeated manner for continued access to resources in a secure environment.

There has been an increasing interest in biometric systems in recent years. Biometrics

technologies are widely used in numerous security applications and are considered among the most accurate and efficient authentication systems on the market. Compared to the conventional methods, biometrics can provide enhanced security and convenience by the use of a person's unique, permanent, and universal features recognized from human characteristics. There are two main types of biometric identifiers: one is physiological characteristics, i.e., the shape or composition of the body, and the other is behavioral characteristics, i.e., the behavior of an individual. Examples of physiological characteristics for biometric authentication are eye position tracking, pupil size, skin conductivity, relative blood flow, fingertips, DNA, face, retina, or ear features. Behavioral characteristics are traits that are learned from human actions, such as typing rhythm, gesture, pace, and voice. Dynamic signature verification and keystroke dynamics are two examples of behavioral characteristics. So far, behavioral biometrics have been less successful compared to physiological ones (Bergadano et al., 2002) because of their significant variability over time.

While some of the above measures may be defined as strong biometrics, others may provide weak biometrics information for many people but can be discriminative for others. An example of such a measure is the relative blood flow. Many people may have similar finger-tip blood flow wave patterns, while some people may have a unique waveform that can be used as a strong identifier. In this paper, we will investigate the use of a finger-tip blood flow waveform as the biometric to perform identity authentication.

1.2 Functional Data

Functional data analysis considers responses and predictors for a subject not as a scalar or vector-valued random variables but instead as random functions defined at infinitely many points. A functional variable $y(t)$ denotes a variable associated with t, i.e., a function of t, where t could be time or some other temporal or spatial variable. It could be multidimensional, although, in most cases, we will limit our discussion to a univariate t. When $y(t)$ is

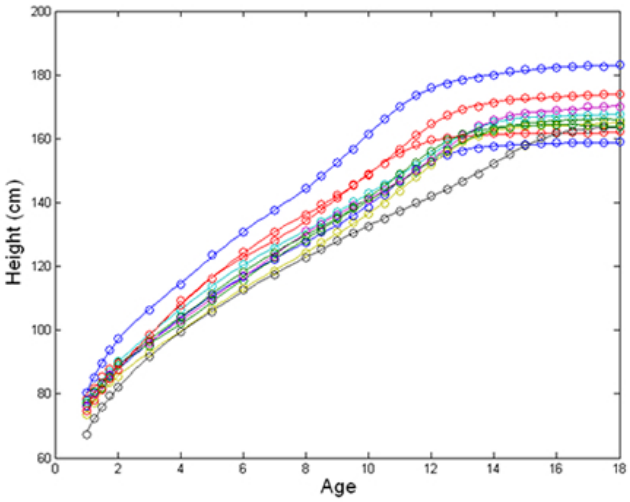

Figure 1.1: *The heights of 10 girls taken between 1 and 18 years. Circles indicate the ages at which measurements were taken. These data were collected as part of the Berkeley Growth Study (Jones and Bayley, 1941).*

continuous, a functional variable is called a curve. More formally, a variable is called a functional variable if it takes values in an infinite dimensional space (Ferraty and Vieu, 2006). Observations of functional variables are called functional data, such as $y(t_i), i = 1, 2, \cdots, n$.

Figure 1.1 shows an example of functional data. It displays the height function measured between 1 and 18 years for each of 10 girls. We view each curve as a functional observation. Circles indicate the ages at which measurements were taken. Growth is the most rapid in the earliest years and slows down after the growth spurt that occurs at ages ranging from about 9 to 15 years. One girl is tall for all ages, but some girls can be tall during childhood, but end up as a comparatively small adult.

In many research areas, such as medicine, biology, economics and engineering, the data generating process is naturally a stochastic function. Moreover, many problems are better approached if the data are considered as functions. For instance, if each curve is observed at different points, a multivariate analysis would not be valid, and it is necessary to smooth the data and treat them as continuous functions defined over a common interval.

In the field of functional data analysis, there are two viewpoints based on how they conceptualize functional data (Ramsay, 1982). On the one hand, some authors view functional data as a smoothed version of the multivariate data, and functional data analysis applies the multivariate data analysis techniques in the language of functional observations. On the other hand, the statistical application of spline functions, especially in the scope of nonparametric function estimation (Eubank, 1999; Silverman et al., 1985; Wahba, 1990) speeds the development of functional data analysis. Although there is a difference between these two thoughts, the fundamental idea is that we view the observation as a whole function defined on a bounded interval, rather than focusing on the individual values at particular points in the range.

In our project, pulse waves are an example of functional data where we view each waveform, instead of individual values at particular points, as a functional observation. Let $y_{ij}(t)$ be the near-infrared PPG signal from the pulse oximeter for the j^{th} curve of person i at time point t, where $i = 1, 2, \cdots, 48$, $j = 1, 2, \cdots n_i$, and n_i is total number of pulse waves collected for person i.

Typically, in functional data analysis, the primary goal is to discover something about the smooth curves that underlie the functional observations and to study the behavior of the entire set of functional data (consisting of many curves). Therefore, smoothing usually is applied before analyzing functional data. Ramsay (2006) also mentioned that the first steps in a functional data analysis are data representation: smoothing and interpolation, data registration or feature alignment, and data display.

Figure 1.2: *A finger pulse oximeter. Figure from SOS Technologies.*

1.3 A motivating example: Authentication by Photoplethysmography

The utilization of photoplethysmographic (PPG) signals for biometric identification represents a novel approach in the area of secure authentication. The word plethysmograph is a combination of two Greek words "plethysmos" which means increase and "graph" which means writing (Shelley, 2007). It is mainly used to determine the blood volume or blood flow in the body associated with each heartbeat. The introduction of the pulse oximeter (Figure 1.2) into routine clinical care greatly emphasizes the importance of PPG. Photoelectric plethysmography, also known as PPG, can estimate the skin blood flow using infrared light. It uses a probe which contains a light source and detector to detect cardiovascular pulse wave that propagates through the body. There are two application modes for a pulse oximeter: transmission and reflective (Figure 1.3), depending on where the detector is placed. In the transmission mode, the detector is placed on a thin part of the body, usually a fingertip or earlobe. The light will pass through the body part to the photodetector. The detector can also be placed next to the light source for reflective plethysmography. Reflective pulse oximetry does not require a thin part of the body and is thus suitable for a universal application such as the forehead, feet, and chest. In outpatient settings, pulse oximeters are

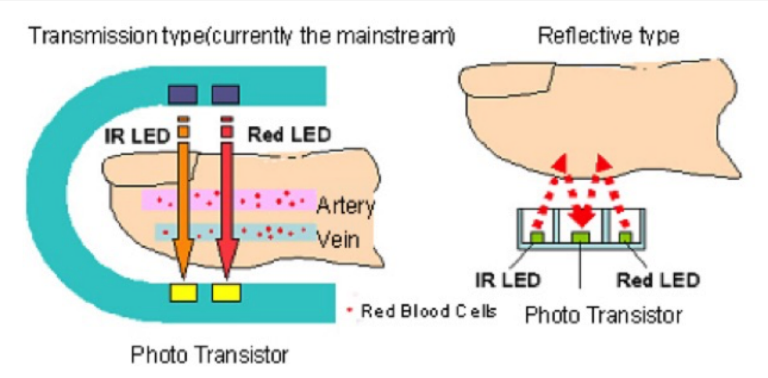

Figure 1.3: *Transmission and reflective type pulse oximeter. Graph from Jian Shao and Qirui Xu, Pulse Oximeter Based Heart Beats Rate Monitor, EE6350 VLSI Design Lab.*

commonly worn on the finger. However, in cases such as shock or hypothermia, blood flow to the periphery can be reduced, resulting in a PPG without a discernible cardiac pulse. In this case, a PPG can be obtained from a pulse oximeter on the head, with the most common sites being the ear, forehead, or nasal septum.

The PPG signal reflects the blood movement in the vessel, which goes from the heart to the fingertips and toes through the blood vessels in a wave-like motion (Tokutaka et al., 2009), as the amount of the backscattered light corresponds to the variation of the blood volume (Alnaeb et al., 2007). Figure 1.4 shows a typical waveform of the PPG and its characteristics features. It is most useful to view the waveform from the pulse oximeter as the measure for the change in blood volume during a heartbeat cycle. The appearance of PPG pulse can be divided into two stages: the anacrotic stage is the rising edge of the waveform, whereas the catacrotic phase is the decreasing edge of the waveform as shown in Figure 1.4. The mounting part is primarily associated with systole and the falling part with diastole. A dicrotic notch, shown in Figure 1.4, usually occurs during the decreasing phase of people with healthy compliant arteries. The systolic amplitude ($\mathbf{x}$ in Figure 1.4) is an indicator of the pulse changes in the blood volume caused by arterial blood flow around the measurement site (Asada et al., 2003; Chua and Heneghan, 2006). An example of four consecutive waveforms are shown in Figure 1.5. Some features based on the PPG have been described

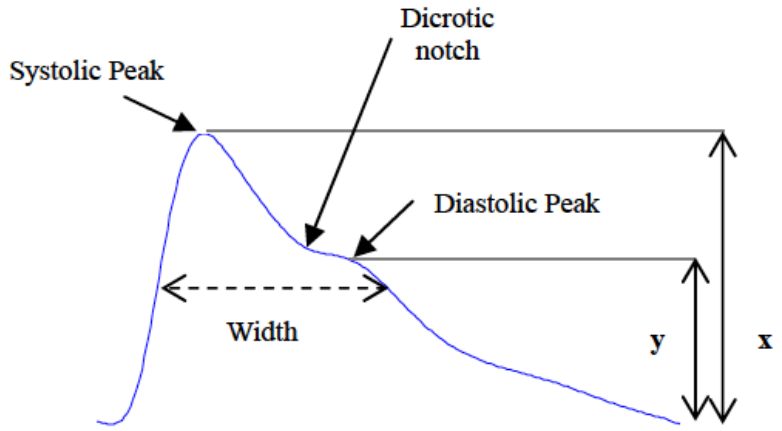

Figure 1.4: *A typical waveform of the PPG and its characteristics parameters: the amplitude of the systolic peaks is x while y is the amplitude of the diastolic peak. Figure from Elgendi (2012).*

in the literature.

1) Systolic Amplitude: As shown in Figure 1.4, the systolic amplitude (**x**) is an indicator of the pulsatile changes in blood volume caused by blood flow around where the measurement was taken (Asada et al., 2003; Chua and Heneghan, 2006). Systolic amplitude has been related to stroke volume (Murray and Foster, 1996), which is the volume of blood pumped from the left ventricle per beat. It is also has been suggested that systolic amplitude is potentially a more suitable measure than the pulse arrival time for estimating continuous blood pressure (Chua et al., 2010).

2) Pulse Area: The pulse area is measured as the total area under the PPG curve. Seitsonen et al. (2005) found that the PPG area respond to skin incision to differ between movers and non-movers.

3) Pulse Interval: The distance between the beginning and the end of the PPG waveform. Poon et al. (2004) suggested that the ratio of pulse interval to its systolic amplitude could

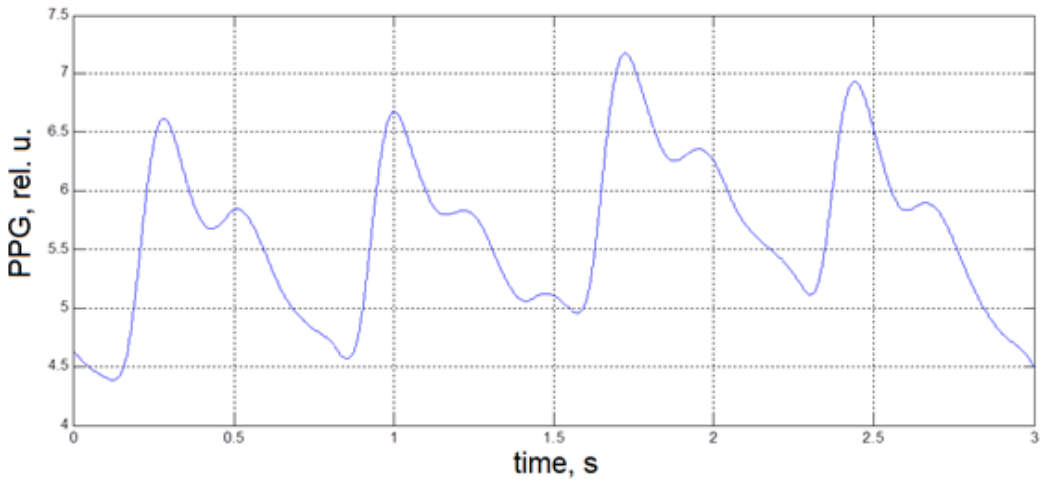

Figure 1.5: *Consecutive PPG curves.*

provide an understanding of the properties of a person's cardiovascular system.

Compared to other biometric approaches, the PPG technique has distinct advantages including low cost, easy to use without any complicated procedure, straightforward but accurate to estimate the skin blood flow, heart rate, and blood oxygen saturation. Moreover, it does not need direct contact with the skin surface, and it is conveniently accessible to the various location of the body, such as the finger, earlobe, wrist or forehead.

1.4 Purpose of the Study

Identifying an individual based on a username, password or other means helps companies ensure that the person is whom he or she claims to be when accessing a system, application or network. However, in some cases, traditional authentication processes are not enough to provide strong security throughout a user work session. That's where continuous authentication comes in. The pulse wave is an important health metric able to perform continuous authentication and also identify people at increased risk for development of heart disease, stroke or some other health issues. The research objective is to differentiate and, more importantly, to identify the subjects using their pulse wave curves from PPG, i.e., identity

assurance. Identity assurance is a set of mechanisms and strategies allowing an organization to minimize the risk related to identifying impersonation and misappropriation of authentication credentials.

1.5 Project and Data Description

Dr. Steve Warren, from the Department of Electrical and Computer Engineering, Kansas State University, provided the pulse wave data for our project. The pulse oximeter prototypes were used to acquire PPG records from 48 different subjects that are 20 to 64 years old. The pulse oximeter was placed on the index finger of each individual, and the adjusted near infrared AC signal was used for data collection. Experimental results were acquired in an indoor environment. Some were collected at Kansas State University Open House, and thus may not be as stable as those gathered in the laboratory. Figure 1.6 illustrate 60 seconds of representative fingertip data from one subject and six single cycles extracted from the raw signal after shifting to the same starting point at zero. Many people may have similar finger-tip blood flow wave patterns, while some people may have quite different waveforms that can be used as a strong identifier.

1.6 Review of Literature

1.6.1 Biometric Identification

Various biometric measures have been investigated for authentication purposes, including face complexion (Brunelli and Poggio, 1993; Samal and Iyengar, 1992), iris recognition (Negin et al., 2000), the electrocardiogram (ECG) (Biel et al., 2001; Odinaka et al., 2012), electroencephalogram (EEG) (Khalifa et al., 2012), phonocardiogram (PCG) (Beritelli and Serrano, 2007), and photoplethysmogram (PPG) (Bao et al., 2005; Elgendi, 2012; Gu and Zhang, 2003; Gu et al., 2003; Spachos et al., 2011; Yao et al., 2007). Monrose and Rubin

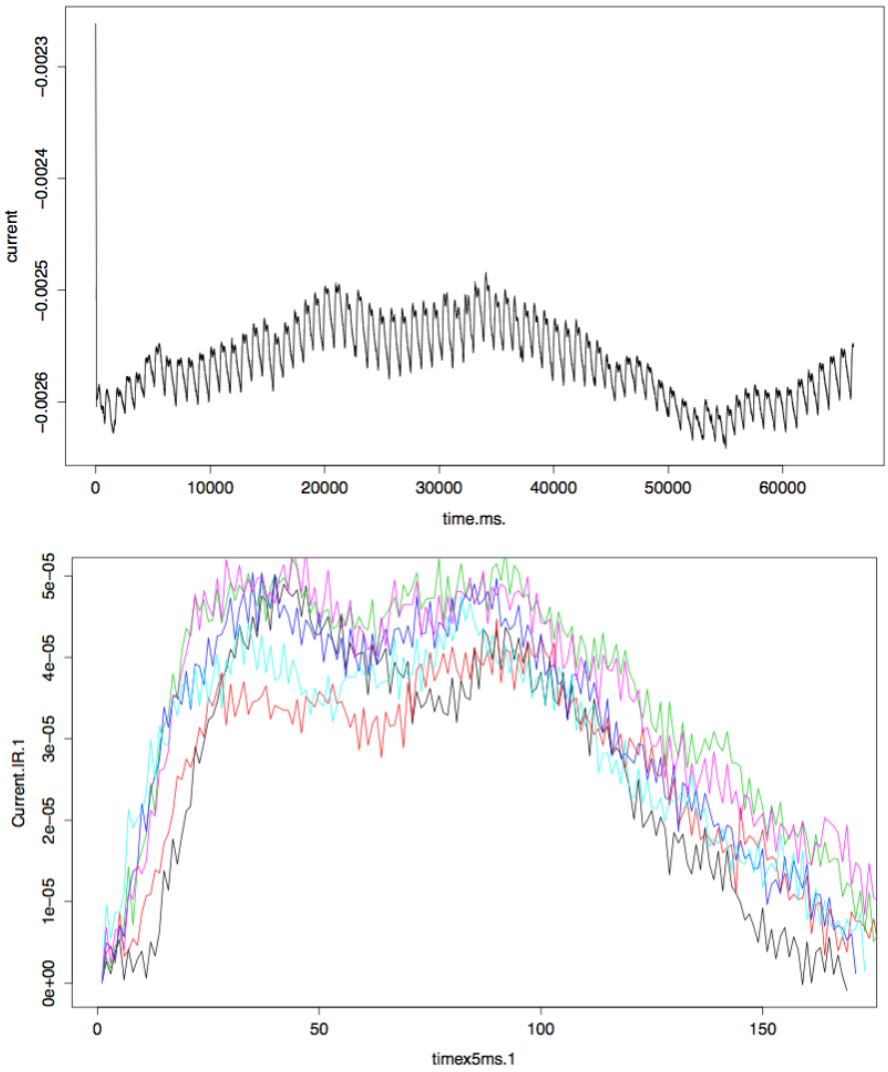

Figure 1.6: *PPG curves collected from a real person in the project. The upper graph is the raw PPG signal within around 60 seconds of time, with the current shown on the vertical axis. The lower one extracted six single cycles from the raw signal and shifted to the same starting point at zero.*

(2000) proposed keystroke biometric technique for continuous authentication. Their method is based on a single biometric (unimodal technique), so in the absence of keystroke data, the system is not able to authenticate the user. Altinok and Turk (2003) proposed continuous authentication techniques using face, voice, and fingerprint. They claimed that a continuous biometric authentication system should be able to provide a meaningful estimate of authentication certainty at any given time, even in the absence of any biometric

data. They presented a new temporal integration technique that satisfied this requirement. Sim et al. (2007) proposed a continuous authentication technique using face and fingerprint biometrics. They used a mouse with a built-in fingerprint sensor, which made fingerprint authentication a passive method for authentication. Derawi et al. (2010) collected data with a commercially available mobile device containing low-grade embedded accelerometers. The mobile device was placed at the hip on each volunteer to collect gait data. Preprocessing, cycle detection and recognition-analysis were applied to the acceleration signal. Sitová et al. (2016) introduced Hand Movement, Orientation, and Grasp (HMOG), a set of behavioral features to continuously authenticate smartphone users. HMOG features capture subtle micro-movement and orientation dynamics resulting from how a user grasps, holds, and taps on the smartphone. Buriro et al. (2016) proposed a mechanism which profiles a user based on how he holds the phone by taking into account the micro-movements of a phone and the movements of the user's finger during writing or signing on the touch screen.

The utilization of PPG signals for biometric identification represents a novel approach in the area of secure authentication. The studies on biometric recognition methods based on PPG signals have been performed on small datasets and by using algorithms inspired by medical analysis techniques. Elgendi (2012) has given a comprehensive analysis of the morphology of fingertip PPG and explained the reasons for their diversity and variation. For most researchers, the primary concentration has been on the use of peak locations, relative amplitudes, derivative-based slopes, and time intervals between the interests points. The method presented in Gu et al. (2003) computes the templates by using four features: the peak number, upward slope, downward slope, and time interval. However, this approach ignores the higher-order derivative information contained in the pulse wave, and therefore, does not make full use of the waveform to improve the identification accuracy. The study presented in Yao et al. (2007) uses values of the local maximum, minimum, and inflection points of PPG pulses to evaluate if samples of genuine and impostor individuals present sufficient discriminability for being used in biometric recognition systems. Kavsaoğlu et al. (2014) have used 40 such features, and they applied a feature ranking algorithm with k-means

clustering. The method described in Spachos et al. (2011) performs biometric recognition by applying the linear discriminant analysis (LDA) and using an eigenspace decomposition of the time-domain signal to obtain a template for identification. These methods work well for datasets with a small number of curves. When the sample datasets get larger, we should develop more appropriate approaches and algorithms which permit performance of biometric recognition in a straightforward and fast manner. To the best of our knowledge, there are no studies on the use of PPG signals in continuous authentication systems or how to standardize the PPG amplitude/rhythm for comparing one waveform to another.

1.6.2 Statistical Methods for Functional Data

Pulse waves are an example of functional data: each waveform, instead of the individual values at particular points, represents a functional observation. Due to its practical advantage, functional data analysis has received considerable attention in diverse areas of applications: the comparison of growth curves (Rao, 1958), the analysis of handwriting in Chinese (Ramsay, 2000), modeling price dynamics in online auctions (Wang et al., 2012), crop lodging assessment (Ogden et al., 2002), forecasting of climate variations (Besse et al., 2000), data mining (Hand, 2007), ozone population forecasting (Damon and Guillas, 2002), and many more. Ramsay and Silverman (2002) give numerous examples considering a continuous functional variable. Often, the continuous functional variable is time, which is the case in our project, even though functional data may be observed over space, wavelength, temperature or other continuums.

Research tools that are useful for handling functional data include various smoothing methods, notably kernel regression, local least squares, and spline smoothing for which various excellent references exist (De Boor et al., 1978; Eubank, 1999; Wand and Jones, 1994), functional analysis (Conway, 2013; Hsing and Eubank, 2015), and stochastic processes (Karlin, 2014). Several software packages are publicly available to analyze functional

data, including software at the Functional Data Analysis website of James Ramsay (http://www.psych.mcgill.ca/misc/fda/), the **fda** package (Ramsay et al., 2015) on the CRAN project of R (R Core Team, 2013), the Matlab package **PACE** on the website of the Statistics Department of the University of California, Davis (http://www.stat.ucdavis.edu/PACE/), and the R package **refund** (Goldsmith et al., 2018) on functional regression.

A considerable effort is being made to adapt some standard statistical methods for functional data. This is the case, for example, of principal component analysis (Boente and Fraiman, 2000; Dauxois et al., 1982; Locantore et al., 1999; Pezzulli and Silverman, 1993; Ramsay, 2006; Silverman et al., 1996), discriminant analysis (Ferraty and Vieu, 2003), regression (Cardot et al., 1999; Cuevas et al., 2002; Ferraty and Vieu, 2002), and analysis of variance (Cuevas et al., 2004; Kaufman et al., 2010). Fan and Lin (1998) proposed a two-sample test and ANOVA test for the mean functions, with further work by Cuevas et al. (2004) and Zhang (2013). Other two sample tests have been proposed for distributions of functional data (Hall and Van Keilegom, 2007) and for covariance functions (Panaretos et al., 2010).

1.6.3 Equivalence Testing

In our project, we are more interested in showing that the pulse waves being compared belong to the person owning the sensor because the mistake of claiming the wrong person is worse than rejecting the correct person. In other words, our inferential goal is to establish practical similarity rather than a statistically significant difference. If we want to claim that two groups of pulse waves are "similar," it is not enough to do the traditional test of difference and just "not reject," since failing to find a difference is not proof of similarity. As a result, we consider equivalence testing where support for the alternative hypothesis of similarity is provided by evidence against the null hypothesis of difference.

There are 3 general approaches for equivalence testing: the *confidence interval approach*, developed by Westlake (1981) and presented in this proposal; the *nonequivalence null hypothesis approach*, developed by Anderson and Hauck (1983), which uses an approximation to a non-central t-distribution to compute the p-value of the test; and *Bayesian methods*, developed by Selwyn et al. (1981) and Selwyn and Hall (1984). The first two methods require the fewest assumptions (Westlake, 1988), and thus, are more appealing. Comparison of the two approaches shows that the confidence interval approach is more conservative since the actual Type I error rate is equal to or less than the stated Type I error rate. For the nonequivalence null hypothesis approach, the actual Type I error rate could be higher than the stated Type I error rate (Anderson and Hauck, 1983).

Equivalence testing for scalar data has been well addressed in the literature. See Berger and Hsu (1996) for a comprehensive overview of commonly used procedures. However, the same cannot be said for functional data. The resultant complexity from maintaining the functional structure of the data rather than using scalar transformation to reduce the dimensionality renders the existing literature on equivalence testing inadequate for desired inference. Alberola-López and Martín-Fernández (2003) discuss a frequentist approach for comparing two functions (time series) through the use of a Fourier basis expansion in the cosinor model. Behseta and Kass (2005) present two methods of testing the hypothesis of equality of two functions in a generalized non-parametric regression framework using a recently developed generalized non-parametric regression method called Bayesian adaptive regression splines (BARS)(DiMatteo et al., 2001). The first method uses Bayes factors, and the second method uses a modified Hotelling-T^2 test. They applied both methods to the analysis of 347 motor cortical neurons. All three approaches test strict equality between the functions of interest but do not establish practical equivalence.

Fogarty and Small (2014) proposed a framework for equivalence testing for functional data with both the frequentist and Bayesian paradigms. Their frequentist hypothesis test extends the Two One-Sided Test (TOST) procedure for equivalence testing to the functional

regime and uses the nonparametric bootstrap (Efron and Tibshirani, 1994) for assessing equivalence by constructing point-wise confidence intervals for the metric of equivalence. Their Bayesian methodology employs a functional analysis of variance model and uses a flexible class of Gaussian processes for modeling the data and the parameters. However, the authors do not explain how they choose the equivalence bands used in their paper, which is the most crucial step in equivalence testing.

Chapter 2

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the practice of using visual and quantitative methods to understand and summarize a dataset, usually to investigate a specific question or to prepare for more advanced modeling, without making any assumptions about its content. It is a critical first step in analyzing the data from an experiment. EDA often relies on visualizing the data to assess patterns and identify data characteristics. It also takes advantage of some quantitative methods to describe the data.

In this chapter, we first apply the functional boxplots in Section 2.1 to detect the outliers and visualize the pulse waveforms. We then implement functional analysis of variance (FANOVA) in Section 2.2 and permutation test in Section 2.3 to differentiate the individuals. Based on these preliminary analyses, we find that the variability among curves between different people is large enough, relative to the variability among curves within a person, to use these curves for authentication. However, the main objective of identity authentication cannot be achieved by either the FANOVA or the permutation test, we propose more appropriate procedures in Chapter 3.

Hitchcock et al. (2007) investigated the possible benefits of pre-smoothing functional data before performing cluster analysis. They compare the accuracy of clustering results by

the use of unsmoothed functional data with two smoothed versions of the data and finds that smoothing produces a more accurate clustering. In our case, we will first apply non-parametric regression smoothing to the pulse waves and then make further analysis, such as functional boxplots, FANOVA, permutation tests and equivalence testing.

2.1 Visualization by Functional Boxplot

To analyze functional data, researchers often used mathematical models, among which Ramsay (2006) provided various parametric methods while Ferraty and Vieu (2006) developed detailed nonparametric techniques. In contrast to model-based analysis, visualization methods often help to display the data, highlight their characteristics, and reveal interesting features.

A traditional boxplot is a robust tool for visualizing univariate data, giving an assessment of the symmetry of the data and identifying outliers. However, this method of assessing univariate data is unable to account for the complexity of functional data where multiple measurements are taken per observation.

When constructing a traditional boxplot, the observations are first ordered by numeric values so that the median observation, the first and third quartiles may be found. Ordering multivariate and functional data is more complicated. For this reason, various methods have been proposed to define a depth score for each multivariate and functional observation by which the observation can be ranked. Examples of depth for multivariate observations include the Mahalanobis depth (Mahalanobis, 1936), the Tukey halfspace location depth (Tukey, 1975), the Oja depth (Oja, 1983), the simplicial depth (Liu et al., 1990), the majority depth (Singh, 1991), and the likelihood depth (Fraiman et al., 1999).

Vardi and Zhang (2000) proposed an L_1-depth which can be extended to functional

observations. Febrero et al. (2007) reviewed the functional depth of trimmed means (Cuevas et al., 2006; Fraiman and Muniz, 2001). López-Pintado and Romo (2009) introduced notions of band depth (BD) and modified band depth (MBD) as ways to order functional data. They allow for ordering a sample of curves from the center outward and, thus, introduce a measure to define functional quantiles and the centrality or outlyingness of a functional observation. Having the order of curves, the functional boxplot is a natural extension of the classical boxplot and is an informative and appealing tool for visualizing functional data. BD is the proportion of a functional observation, i.e., a curve, that falls within a band delimited by two functional observations. Considering a data set with N functional observations, there are $\binom{N}{2} = \frac{N(N-1)}{2}$ possible number of bands catching a single functional observation. BD for a single curve is the ratio of the sum of the number of bands containing this target curve to the total number of possible bands.

Figure 2.1 shows a simple example with four curves on how to compute BD and MBD in practice. First, there are six possible bands delimited by two curves. For example, the gray area in Figure 2.1 is the band delimited by $y_1(t)$ and $y_3(t)$. We notice that the curve $y_2(t)$ falls completely within the band, while $y_4(t)$ only partly does. We define that a curve is contained in a band if this curve is on the border of the band. Then the band depth for the curve $y_2(t)$ is $5/6 = 0.83$ since only the band delimited by $y_3(t)$ and $y_4(t)$ does not completely contain the curve $y_2(t)$ and the band depth for the curve $y_4(t)$ is $3/6 = 0.5$ as it only falls completely within the band delimited by itself and another curve. Similarly, we could compute $\text{BD}(y_1) = 0.5$ and $\text{BD}(y_3) = 0.5$. To compute MBD, we see that the curve $y_2(t)$ is always contained in the five bands, hence $\text{MBD}(y_2) = 0.83$, the same value as BD. In contrast, the curve $y_4(t)$ only belongs to the band in gray 40% of the time, thus $\text{MBD}(y_4) = (3 + 0.4 + 0.4)/6 = 0.63$ by definition. The median curve is the one with the largest BD or MBD score. Sun and Genton (2012) used both the BD and MBD to create functional boxplot.

In the classical boxplot, the box itself represents the middle 50% of the data. While

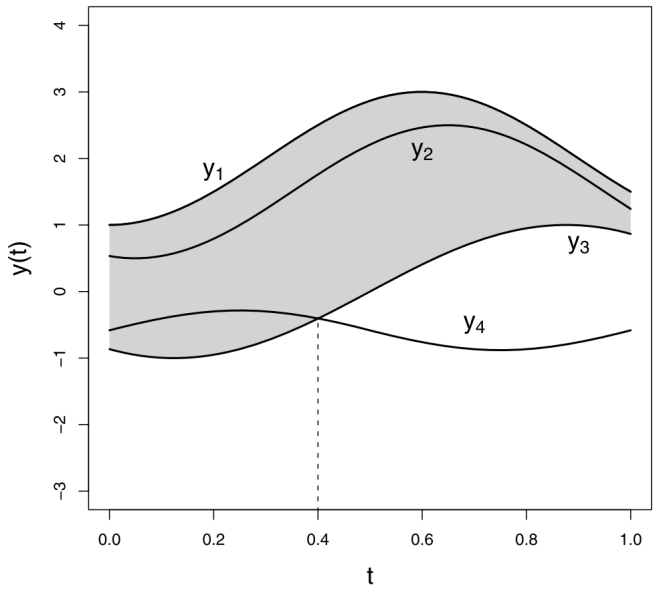

Figure 2.1: *An example of BD and MBD computation: the gray area is the band delimited by $y_1(t)$ and $y_3(t)$. The curve $y_2(t)$ completely belongs to the band, but $y_4(t)$ only partly does. Graph from Sun and Genton (2012).*

in the functional boxplot, the band delimited by the α proportion ($0 < \alpha < 1$) of deepest curves from the sample is used to estimate the α central region. In particular, the sample 50% central region is

$$C_{0.5} = \{(t, y(t)) : \min_{r=1,\cdots,\lceil n/2 \rceil} y_{[r]}(t) \leq y(t) \leq \max_{r=1,\cdots,\lceil n/2 \rceil} y_{[r]}(t)\},$$

where $\lceil n/2 \rceil$ is the smallest integer not less than $n/2$, and $y_{[i]}(t)$ denotes the sample curve associated with the ith largest band depth value at time t. We view $y_{[1]}(t), \cdots, y_{[n]}(t)$ as order statistics, with $y_{[1]}(t)$ being the deepest (most central) curve or simply the median curve, and $y_{[n]}(t)$ being the most outlying curve. The border of the 50% central region is defined as the envelope representing the box in a classical boxplot. Thus, this 50% central region is analogous to the "interquartile range" (IQR) and gives a useful indication of the spread of the central 50% of the curves. We extend the 1.5 times IQR empirical outlier criterion to the functional boxplot. The fences are obtained by inflating the envelope of the 50% central region by 1.5 times the range of the 50% central region. Any curves outside the

fences are flagged as potential outliers.

We made functional boxplots for the waveforms of each person to get an overall idea of the behavior of curves. One issue regarding the pulse waves is that several waveforms do not seem to start at the right place due to the difficulty in the determination of minimum local points and separating individual cycles during the data preprocessing. We also used functional boxplots to edit the curves: identifying the proper starting points of each curve. However, to make the boxplots for the curves, the data format needs to be a $p \times n$ matrix where n is the number of curves, and p is the number of evaluation points for each curve. However, the length of each pulse wave in our dataset varies from 0.6 seconds to 1.1 seconds. Therefore, we first applied the kernel smoothing to the full-length curves, and then truncated the smoothed curves to get equal-length pulse waves. After that, the functional boxplots can be obtained using the **fbplot** function in R package **fda** (Ramsay et al., 2015). Figure 2.2 shows the functional boxplots for person 1, person 2, person 7, and person 10. The dark-colored areas represent the 50% central regions for each of the four persons with the median curves shown in black. The red dashed lines indicate possible outliers. There are three outliers (curves 16, 33, 54) for person 1; one outlier (curve 83) for person 2; three outliers (curves 4, 5, 88) for person 7; four outliers (curves 1, 2, 5, 54) for person 10.

We noticed that, for example, curve 83 of person 2 was an outlier. However, it seemed to start at an inappropriate place and the beginning of the curve may be the tail of the previous one. If we shift this outlier to the left and make it start at a more reasonable point, it may fall within the "interquartile range" and no longer be an outlier. We found the most appropriate beginning and end points of each outlier or improperly behaved curves for person 1, person 2, person 7 and person 10. Figure 2.3 shows the corresponding functional boxplots using the smoothed and shifted data for these four people. We observed that curve 83, after moving to a more proper beginning point, was no longer an outlier. We identified the outlier curves from the functional boxplots and cleaned the data by shifting outliers to a more proper starting point. From Figure 2.3, we can see that these four persons have similar

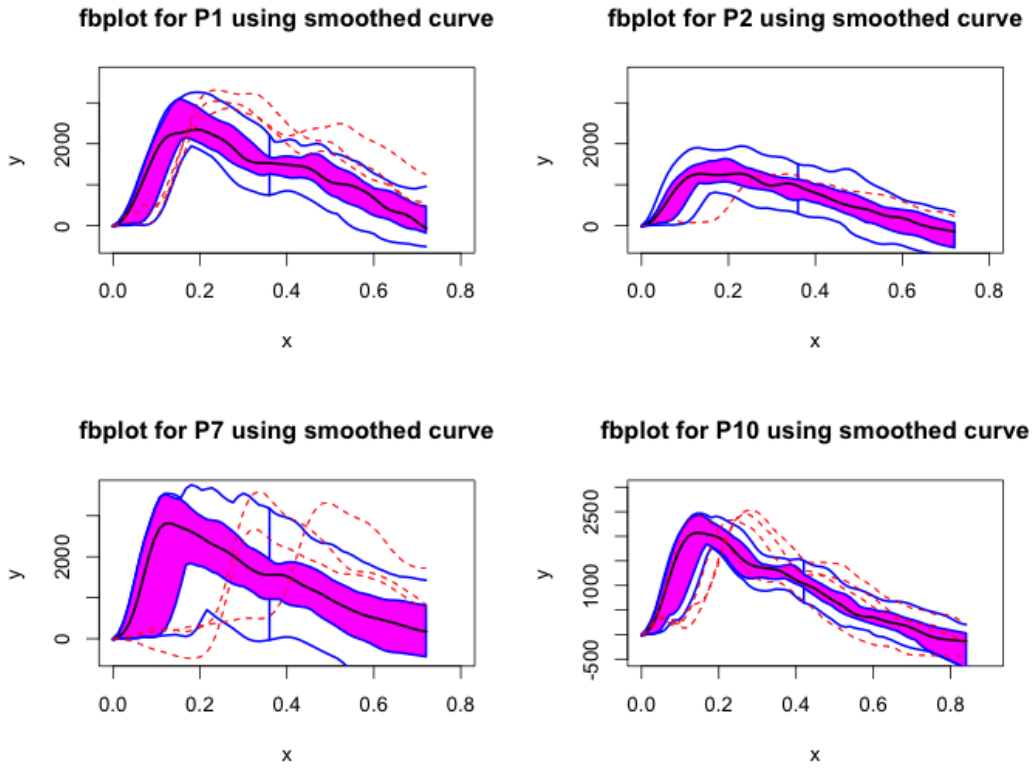

Figure 2.2: *Functional boxplots for person 1, 2, 7, and 10. The colored area represents the 50% central region, with the median curve shown in black. The red dashed lines indicate possible outliers.*

overall pulse wave patterns; however, person 1 and person 10 have a more apparent dicrotic notch than person 2 and person 7. Person 2 has a much lower systolic amplitude and thus was less steep than the other three.

Since the lengths of the curves vary, we also rescaled the time variable on the horizontal axis to a (0, 1) scale so that each curve has the same length. The formula used is

$$t' = \frac{t - \min t}{\max t - \min t}$$

where t is the original value and t' is the normalized value. Figure 2.4 (a) shows the functional boxplot for Person 1 after smoothing and shifting on original curves, and (b) represents the functional boxplot for Person 1 after smoothing and shifting on curves rescaled to (0, 1) on

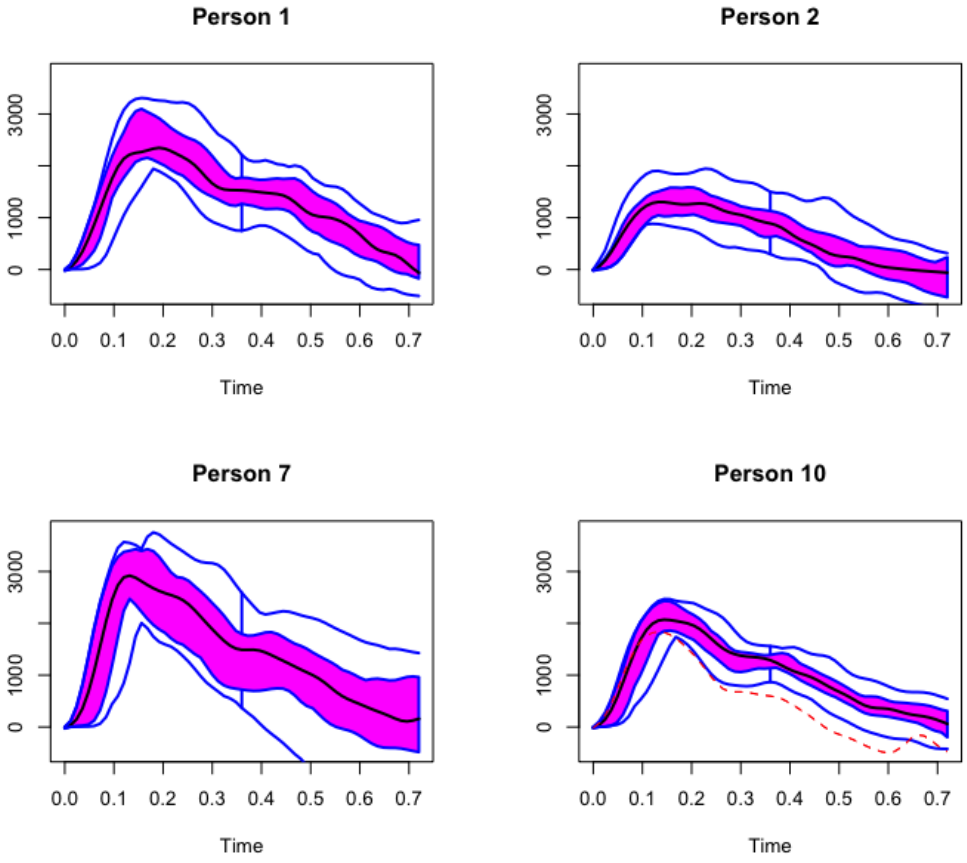

Figure 2.3: *Functional boxplots for person 1, 2, 7, and 10 using smoothed and shifted data.*

time scale. The overall shape of the functional boxplot does not change much. Also note that, for the rescaled data, we can compare relative locations of various features such as the systolic peak or dicrotic notch across subjects.

2.2 Functional ANOVA to Differentiate Persons

Analysis of variance (ANOVA) is a statistical technique that examines differences in the means of a variable across groups of observations. Frequently, we use ANOVA to test for differences among several means by comparing variability among groups relative to variability within groups. We reject the null hypothesis of equality if the between-subject variability is much greater than the within-subject variability. For functional data, the problem of test-

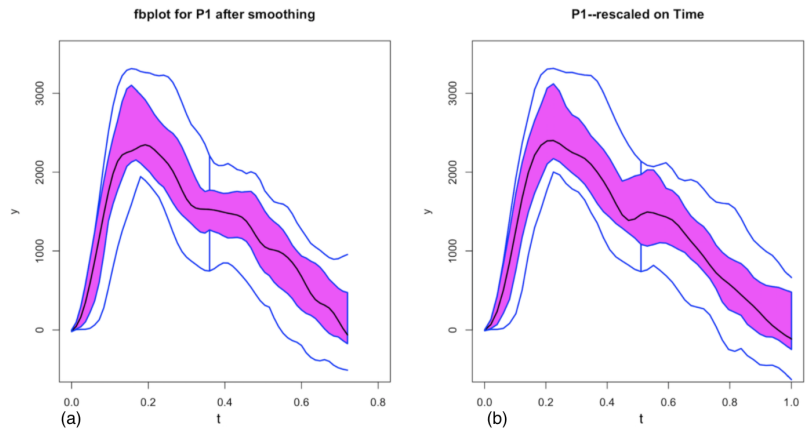

Figure 2.4: *Functional boxplots for person 1 after smoothing. (a): smoothing on original curves; (b): smoothing on curves rescaled to (0, 1) on time scale.*

ing the null hypothesis of equality of their respective mean functions given k independent samples is considered. Functional ANOVA models partition a functional response according to the main effects and interactions of various factors and are appropriate when the data consists of functions that are expected to differ according to some set of categorical factors (Ramsay, 2006). Thus, the setting is quite similar to that of classical one-way ANOVA model except for the k samples under study consist of functional data.

Let $Y_{ij}(t), j = 1, 2, \cdots, n_i, t \in [a, b]$ be k independent samples from L^2-processes $Y_i, i = 1, 2, \cdots, k$, such that $E(Y_{ij}(t)) = f_i(t)$. We want to test the null hypothesis of

$$H_0 : f_1(\cdot) = f_2(\cdot) = \cdots = f_k(\cdot).$$

Fan and Lin (1998) considered the case when the sampling information is in a "discrete" format $Y_{ij}(t), t = 1, \cdots, T$. They proposed a HANOVA (high dimensional ANOVA) test which relies on wavelet thresholding techniques. Maldonado et al. (2002) used a method based on permutation distributions and provided an example in neurophysiology. Kaufman et al. (2010) developed a general framework for functional ANOVA modeling from a Bayesian viewpoint assuming Gaussian Process prior distributions for each batch of functional effects.

They further assumed that the covariance between errors can be specified as a member of Matérn covariance functions.

Cuevas et al. (2004) proposed a simple natural test for one-way ANOVA model for functional data. The test procedure can be seen as an asymptotic version of the well-known F-test, and the test statistic is given by

$$V_n = \sum_{i<j} n_i ||\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}||^2,$$

where $\bar{Y}_{i\cdot} = \bar{Y}_{i\cdot}(t) = \sum_{j=1}^{n_i} \frac{Y_{ij}(t)}{n_i}$, $Y_{ij}(t)$ represents the observed value from the j^{th} curve of the i^{th} subject at time t, n_i is the number of functional observations for subject i, and $||\cdot||$ stands for the usual L^2 norm, $||y|| = \sqrt{\int_a^b y^2(t)dt}$.

The asymptotic distribution of V_n under H_0 is linear combinations of independent Gaussian processes. Provided that n_i are large enough, null hypothesis is rejected at a level of α, whenever $V_n > V_\alpha$, where $P_{H_0}(V_n > V_\alpha) = \alpha$. Then an asymptotic Monte Carlo procedure is implemented to approximately evaluate V_α. The test can be easily carried out using the function **anova.onefactor** in R package **fda.usc** (Bande and de la Fuente, 2016). Figure 2.5 and 2.6 show the functional ANOVA based on 1000 bootstrap resamples for comparing Person 1, 2, 7, 10, 11, and 14 using the truncated and rescaled curves after smoothing and shifting the outliers. The p-values for comparison among six persons using either the truncated or rescaled curves are 0's, indicating that at least one person has different underlying waveform from the others. Person 1 (red) behaves like Person 7 (blue) regarding functional mean, so we performed the functional ANOVA comparing these two individuals in Figure 2.7 after shifting and truncating. There are 15 pairwise comparisons for a small group of six persons. Therefore, with the Bonferroni correction, we test each individual hypothesis at a significance level of $\alpha/15$, where α is the desired overall significance level. The test statistic for comparing between Person 1 and Person 7 is 1063.419 with a p-value of 0.061 after the

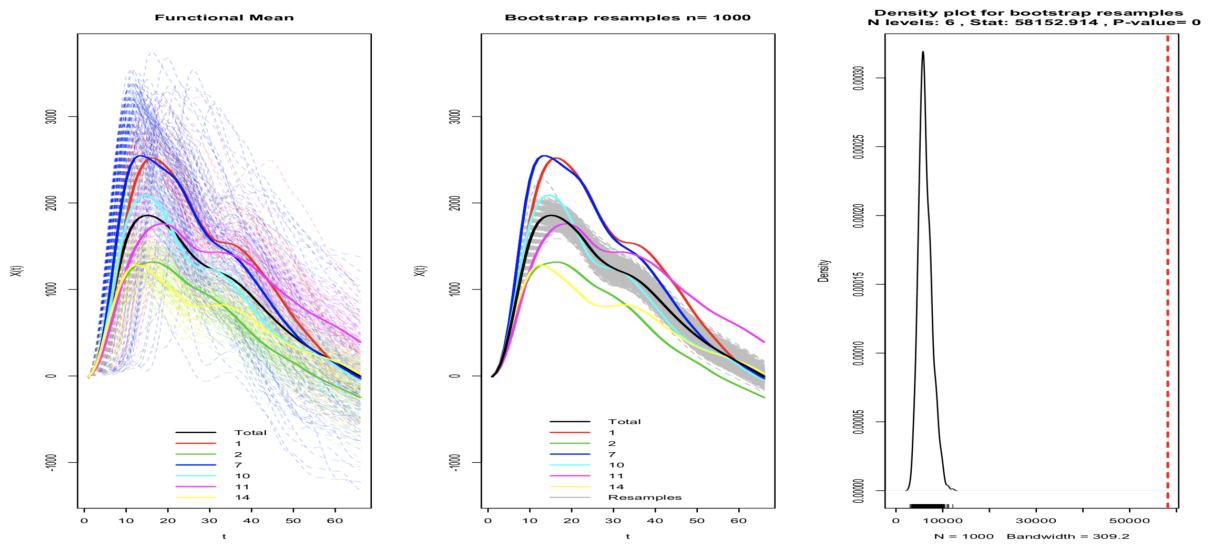

Figure 2.5: *Functional analysis of variance for six persons. A p-value of 0 indicates a significant difference among 6 individuals.*

Bonferroni correction. Therefore, we do not have enough evidence to differentiate these two people from their underlying waveforms at 5% overall significance level.

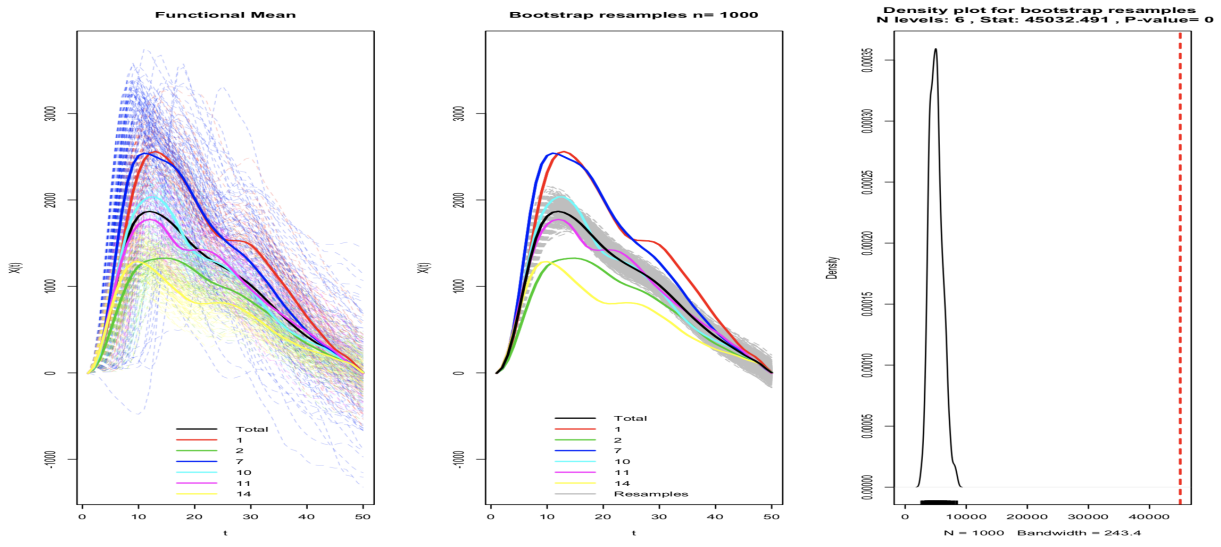

Figure 2.6: *Functional analysis of variance for six persons using data rescaled to (0, 1) on time scale. A p-value of 0 indicates a significant difference among 6 individuals.*

We also performed functional ANOVA on another small group of four people (Person 3, 6,

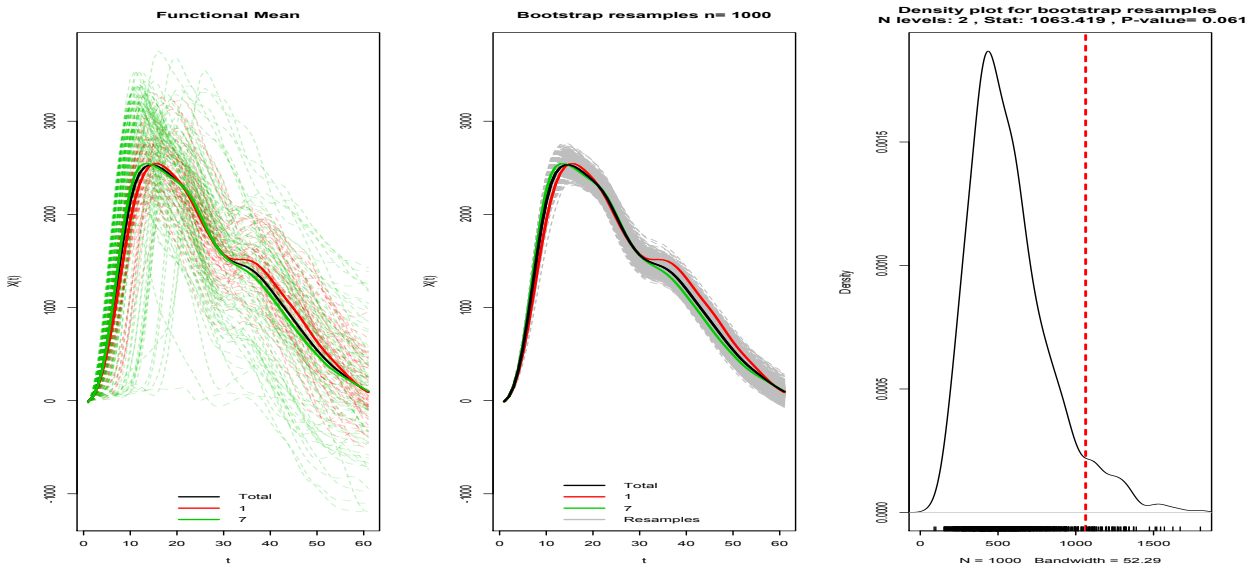

Figure 2.7: *Functional analysis of variance for comparing Person 1 and Person 7. A p-value of 0.061 shows that we do not have enough evidence to differentiate these two people.*

8, and 12) in the sample, and significantly different underlying waveforms could be detected for all the comparisons we made. Therefore, we feel reasonably confident that the variability among curves of different people is great enough, relative to the variability of curves within people, to use these curves for authentication.

2.3 Permutation Test for Functional Data

Another way in which we assess the appropriateness of this data for authentication is permutation tests, an approach for carrying out non-parametric tests. Permutation tests are most useful when we have insufficient information about that distribution of the response variable, are uncomfortable making assumptions about the distribution, or if the distribution of the test statistic is not easily computed or derived. In our case, pulse wave data itself is complex, and the distribution of the test statistic is difficult to derive. Thus, the permutation test is an option to have a general idea on the behavior of the waveforms.

In our case, we will employ a permutation test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data. In contrast to many popular "classical" statistical tests, such as the t-test, F-test, z-test, and χ^2-test, whose reference distributions are obtained from theoretical probability distributions, permutation tests build the reference distribution by resampling the observed data. Like bootstrapping, a permutation test does not rely on traditional distribution assumptions and the only assumption required is that the samples are assumed to be independent.

In each iteration of a permutation test, we rearrange the data set and compute the test statistic from the shuffled data. The proportion of the shuffled test statistics that are as extreme and more extreme than the observed test statistic gives a p-value. While a permutation test requires that we perform all possible permutations of the data, which can become quite large, we can efficiently conduct approximate permutation tests by conducting a large number of random resamples. That process should, in expectation, approximate the permutation distribution.

For our project, we first performed the permutation test on four persons (person 1, 2, 7 and 10) to see if they share the same population mean waveform. There are 57 curves for person 1, 89 curves for person 2, 92 curves for person 7, and 56 curves for person 10. The null and alternative hypotheses for the test can be stated as follows:

H_0 : Four persons have identical mean waveform.

H_a : At least one person has different mean waveform from the others.

We selected the middle 40 curves from each of the four persons since curves in the middle part of the data collection process were more stable and reliable than those from the beginning and the tail. A nonparametric regression estimate using local polynomial kernel GEE method (Chapter 4 of Wu and Zhang (2006)) was obtained by applying kernel smoothing to

the full-length curves before making further analysis.

We consider the area between two estimated curves as our test statistic, and we compute it from the trapezoidal rule. The trapezoidal rule is a technique for approximating the definite integral $\int_a^b f(t)dt$. The trapezoid rule works by approximating the region under the graph of the function $f(t)$ as many trapezoids and calculating the sum of their areas. Let t_k be a partition of $[a, b]$ such that $a = t_0 < t_1 < t_2 < \cdots < t_{N-1} < t_N = b$ and Δt_k be the length of the k^{th} subinterval, that is, $\Delta t_k = t_k - t_{k-1}$. Then

$$\int_a^b f(t)dt \approx T_N = \sum_{k=1}^{N} \frac{f(t_{k-1}) + f(t_k)}{2} \Delta t_k. \tag{2.1}$$

Since the time points are equally-spaced after smoothing, $\Delta t_k = \Delta t$ is a constant. The approximation becomes more accurate as the resolution of the partition N increases.

Let $C_1(t), C_2(t), C_7(t)$ and $C_{10}(t)$ represent the local linear regression estimates from each person, and $\tilde{C}(t)$ is the estimated curve using local linear regression based on all 120 curves. Let $f_i(t_k) = C_i(t_k) - \tilde{C}(t_k)$, where $i = 1, 2, 7, 10$. Therefore, the test statistic can be written as

$$\begin{aligned} T &= \frac{\Delta t}{2} \sum_i \sum_k |f_i(t_{k-1})| + |f_i(t_k)| \\ &= 790.824 \end{aligned}$$

We performed 1000 random rearrangements or shuffles in the permutation test, and none of the test statistics from permutation is larger than the observed test statistic $T = 790.824$. Thus, the probability of obtaining results as or more extreme than the observed test statistic T in this set is zero. In another words, the permutation p-value is approximately 0. Thus, at least one person has a different underlying waveform from the others at 5% significance level.

P-value	Person 1	Person 2	Person 7	Person 10
Person 1	–	0	0.002	0.001
Person 2		–	0	0
Person 7			–	0
Person 10				–

Table 2.1: *Permutation p-values for comparing between two individuals using a Bonferroni adjustment.*

Next, we continue to make pairwise comparisons. The permutation p-values for pairwise comparisons among Person 1, 2, 7, and 10 using a Bonferroni adjustment are shown in Table 2.1. All of the p-values are significant at 5% significance level. Thus, we can successfully differentiate between any pair among these four people.

Pulse waves collected at a different times of the day from the same person may not have an identical mean pattern. Permutation tests can also be used to test if curves from the same individual have different underlying waveforms. We consider smoothed and shifted data for the permutation test for person 1. The first 30 curves are grouped as the training data and the rest 27 curves as the test data. The area computed from trapezoidal rule between the two estimated curves from training and test set is 46.758. We used 1000 random permutations and computed the area between two groups from the trapezoidal rule. We found that 530 test statistics out of 1000 are bigger than 46.75847. Thus, the two-sided p-value is 0.53. The p-value for comparison within Person 1 using curves rescaled to (0, 1) on timescale is 0.566. At 5% significance level, there is no evidence to conclude that training and test from Person 1, using either original pulse waves or waves rescaled to (0, 1) on timescale, have different probability distributions. We performed the same procedure on Person 2, 7, and 10. The results are shown in Table 2.2. We obtained non-significant results for within-subject comparisons on Person 1, 7, and 10. That is to say, if we have two groups of pulse waves from each of these three people, we do not have enough evidence to distinguish between training and test sets. At 5% significance level, the p-value for comparison within person 2 is significant. One of the possible reasons is that the pulse wave data is collected at K-

Within a person	Original curves	Rescale on time
Person 1	0.53	0.566
Person 2	0.001	0.05
Person 7	0.32	0.388
Person 10	0.187	0.156

Table 2.2: *Permutation p-values for comparing within a person using original curves, and curves rescaled to (0, 1) on time scale.*

state Open House. Thus, subject may have been excited and not quite calmed down at the beginning of the data collecting process resulting in different behavior occurred between the early and late parts of the waveforms.

2.4 Summary

In this chapter, we visualized and explored the pulse waves by looking at the functional boxplots. We also used functional boxplots to correct outliers and shift them back to a more proper starting point. FANOVA and permutation test can quickly and accurately differentiate people if the pulse waves do not belong to the same person. However, the primary goal of identity authentication is still not achieved due to the nature of traditional hypothesis testing because a non-significant test for the difference is not the same as the proof of similarity. In the next chapter, we introduce equivalence testing that could be used to confirm a person's identity, i.e. the pulse waves are actually from the same person. An equivalence test aims to establish equivalence by rejecting a null hypothesis which states that a meaningful difference exists. It is a statistical test of hypotheses for which the inferential goal is to provide evidence of practical equivalence rather than declare a significant difference.

Chapter 3

Equivalence Testing for Identity Authentication

From Chapter 2, we are reasonably confident that the variability among curves between different people is large enough, relative to the variability among curves within a person, to use these curves for authentication. However, the goal of identity authentication is still not achieved by either the FANOVA or the permutation test. In this chapter, we employ the content of equivalence testing and propose a new testing procedure to confirm the identity.

We first give an introduction to equivalence testing in Section 3.1.1 and then explain why we have chosen an equivalence test for identity authentication in Section 3.1.2. Equivalence testing is commonly used in the approval process for generic drugs; thus in Section 3.2 we introduce the assessment of bioequivalence (BE) of two drugs using the two one-sided test procedure (TOST), the most common statistical method for validating equivalence. We also review guidelines for choosing the equivalence bands (EB) suggested by FDA. In Section 3.3.2, we propose a new testing procedure, the nonparametric bootstrap functional equivalence test (NBFET) which could be implemented to evaluate equivalence for functional data.

3.1 Equivalence Test

Ideally, scientists should be able to provide evidence for the absence of a meaningful effect. Currently, researchers often conclude an effect is absent based on a nonsignificant test for difference. Unfortunately, the logic of this approach is flawed; non-significance does not provide evidence that the null hypothesis is true. A widely recommended alternative within a frequentist framework is to test for equivalence. An equivalence test aims to establish equivalence by statistically rejecting a null hypothesis which states that a meaningful difference exists. Equivalence testing originates from the field of pharmacokinetics (Hauck and Anderson, 1984), where researchers sometimes want to show that a generic version of an existing drug or "pioneer" has, to a practical extent, the same pharmacokinetic properties as the pioneer; therefore, the generic drug can be prescribed to patients in place of the pioneer drug.

3.1.1 What is an equivalence test?

An equivalence test is a statistical test of hypotheses for which the inferential goal is to establish practical equivalence rather than declare a statistically significant difference (Berger and Hsu, 1996). For example, we can use an equivalence test to determine whether the means for two groups of measurements are close enough to be considered equivalent. The term "equivalence" is employed here to denote a weaker, or fuzzy, form of an identity relation. The fuzziness of the equivalence hypotheses comes from the fact that exact equality is not required to establish equivalence. Rather, the hypothesis of zero difference, corresponding to the null hypothesis of the traditional two-sided testing problem, is expanded into an interval or "indifference zone" (Wellek, 2010). Differences between two treatments small enough to fall into this interval are deemed equivalent for practical purposes. For example, a deviation of 1 or 2 mg from a targeted 200 mg dose of a drug may be unlikely to have any practical effect when taken by a patient.

A crucial first step in performing a test of equivalence is deciding which parameters to

compare and then defining how large the difference between them must be for the difference to be considered practically important. For example, in the case of the simple parallel group design with two treatment arms, we may want to test whether the population means of the reference group and the test group are close enough to be considered equivalent. For simplicity, assume that the two groups have the same variance, i.e., $\sigma_T^2 = \sigma_R^2 = \sigma^2$. A test of equivalence may be based on the difference of the two means, $\mu_T - \mu_R$. Equivalence of the two groups could be concluded if we can demonstrate that the difference between the two means is small, i.e.,

$$-\epsilon_l < \mu_T - \mu_R < \epsilon_u,$$

where ϵ_l and ϵ_u are some small non-negative known constants. The interval $(-\epsilon_l, \epsilon_u)$ would define our zone of indifference, and the values $-\epsilon_l$ and ϵ_u are then referred to as equivalence limits. The actual, numerical values of the equivalence limits must be assigned *a priori*, i.e., without knowledge of the data. As the guidance for bioequivalence studies, the United State Food and Drug Administration (FDA) recommended using the specifications $\pm 20\%$ of the reference mean as the equivalence limits.

Examples of scenarios requiring equivalence tests include the assessment of a generic drug performance relative to a brand-name drug and a comparison study in which the agreement of new measurement device with the "gold standard" for measuring a particular phenomenon must be assured before the new device can replace the old one. Recently, equivalence testing has been growing in use in scientific research outside its traditional role in the drug approval process and has made inroads in scientific applications unrelated to drug development (Barnett et al., 2006, 2007). Barker et al. (2002) propose to use equivalence test for measuring disparities in vaccination coverage, and Tempelman (2004) applies bioequivalence hypothesis testing to dairy nutrition studies. Parkhurst (2001) states that "equivalence tests improve the logic of significance testing when demonstrating similarity is important."

3.1.2 Why use an equivalence test?

In the usual two-sample two-sided hypothesis test, also called a test of difference, the null hypothesis is that the two experimental groups under study do not differ with respect to a particular parameter, e.g., the mean. If the analysis reveals a statistically significant difference between groups, the null hypothesis of no difference is rejected, and the two groups are then declared to be different. If the analysis fails to indicate that a statistically significant difference between groups exists, the null hypothesis cannot be rejected. However, failing to find a difference is not the same as a proof of similarity. Non-significance may be due to having a sample size that is too small to detect the true difference. The result of non-significance has a history being misinterpreted in some research (Farrington and Manning, 1990; Huh, 1994; Westlake, 1981). Psychologists have incorrectly concluded there is no effect based on a non-significant test result. For example, the word "no effect" has been used in 108 articles published in Social Psychological and Personality Science up to August 2016. Manual inspection revealed that the conclusion of "no effect" was based on statistical non-significance in almost all of these articles. Finch et al. (2001) reported that in the Journal of Applied Psychology, an average of around 38% of articles with nonsignificant results makes a conclusion that accepts the null hypothesis. This practice is therefore problematic. If the researcher truly wants to be able to claim that the two groups are statistically "similar," it is not enough to perform a two-sided test of difference and "not reject" the null hypothesis since failing to reject a null hypothesis of no difference is not logically equivalent to providing evidence for a claim equivalence. As Altman and Bland (1995) put it: "absence of evidence is not evidence of absence."

One critical objective in the present work is to show that the pulse waves being tested do in fact belong to the person owning the sensor. Because the mistake of claiming the wrong person is worse than failing to identify the correct person, it is important to determine whether a person's curves are unique enough to be used for authentication. In other words, our inferential goal is to establish practical similarity rather than detect a statisti-

cally significant difference. Therefore, we consider a test of equivalence where the alternative hypothesis of similarity is supported by rejecting the null hypothesis of difference.

3.2 The Assessment of Bioequivalence

Two different formulations of the same drug are said to be bioequivalent if they are absorbed into the blood and become available at the drug's site of action at about the same rate and concentration. The determination of bioequivalence (BE) is crucial because the approval of generic drugs in the United States often requires the establishment of bioequivalence between the name brand drug and the proposed generic drug. If the manufacturer of the generic drug can demonstrate BE, it can avoid performing costly clinical trials to prove the safety and efficiency of its generic drug.

In many bioequivalence studies, the concentration of the active ingredient of interest is measured in the blood plasma or serum over time. These concentration/time measurements can be connected with a curve, and several variables can be observed. Figure 3.1 shows an example of the concentration-time profile of two drugs. Instead of analyzing the individual profile as multivariate or functional data, the profile is usually reduced to a small number of real values by computing one or more of the following pharmacokinetic characteristics:

- area under the curve as a whole (AUC),
- maximum concentration (C_{max}), and
- time at which the peak concentration C_{max} was measured (t_{max}).

The two drugs are bioequivalent if the population means of these variables are sufficiently close. Commonly, AUC and C_{max} are considered as alternative estimates of the extent of the absorption process for a trial subject, whereas t_{max} is interpreted as a measure of rate. All three parameters are considered as a reasonable option for a reference measure of bioe-

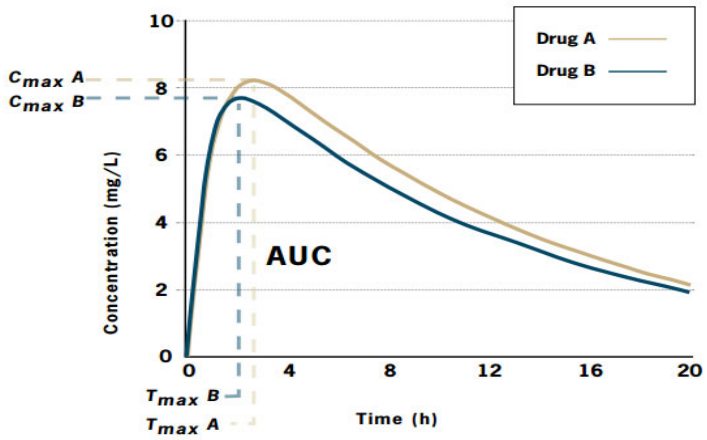

Figure 3.1: *Example of concentration-time profiles for drug A and drug B. Birkett (2003).*

quivalence.

For example, let μ_T denote the population mean AUC for the generic (Test) drug and μ_R denote the population mean AUC for the name brand (Reference) drug. To demonstrate bioequivalence, the following hypotheses are tested:

$$H_0: \quad \frac{\mu_T}{\mu_R} \geq \theta_U \ \ or \ \frac{\mu_T}{\mu_R} \leq \theta_L$$
$$H_a: \quad \theta_L < \frac{\mu_T}{\mu_R} < \theta_U.$$

The values of θ_L and θ_U are specified by the researchers and define how "close" the versions of the drug must be to be declared bioequivalent. Here, the interval (θ_L, θ_U) presents the zone of indifference. Currently, both FDA (1992) and the European Community (EC, 1993) use $\theta_L = 0.8$ and $\theta_U = 1.25$, values that are symmetric in the ratio scale, i.e. $1/0.8 = 1.25$.

Commonly, natural logarithms of the AUC values are computed before analysis so that

more appropriate hypotheses are:

$$H_0: \quad \eta_T - \eta_R \geq \epsilon_U \ \ or \ \eta_T - \eta_R \leq \epsilon_L$$

$$H_a: \quad \epsilon_L < \eta_T - \eta_R < \epsilon_R$$

where $\eta_T = \log(\mu_T)$, $\eta_R = \log(\mu_R)$, $\epsilon_L = \log(\theta_L)$, $\epsilon_U = \log(\theta_U)$. With $\theta_L = 0.8$, $\theta_U = 1.25$, $\epsilon_U = -\epsilon_L$, the standards are symmetric. The Type I error rate is the probability of claiming the drugs to be bioequivalent, when in fact they are not. By setting up the hypotheses above and controlling the Type I error rate at a specified small value, say, $\alpha = 0.05$, the consumer's risk is being controlled.

3.2.1 Equivalence Tests for Differences Between Two Independent Means

The formulation of equivalence hypotheses leads to the most basic form of equivalence testing, the two one-sided tests (TOST) procedure (Barker et al., 2001; Berger and Hsu, 1996; Huh, 1994; Schuirmann, 1987). Largely due to its ease of use and recommendations from FDA guidance, the TOST procedure is the most common statistical method for testing equivalence. The goal in the TOST approach is to specify a lower and upper bound, such that the results falling within this ranges are declared equivalent. As the name implies, the TOST is a two-step procedure and consists of decomposing the interval hypothesis H_0 and H_a above into two sets of one-sided hypotheses. For example, the hypotheses of equivalence testing for differences between two independent means are set up as follows:

$$H_{01}: \quad \eta_T - \eta_R \geq \epsilon_U$$

$$H_{a1}: \quad \eta_T - \eta_R < \epsilon_U,$$

and

$$H_{02}: \quad \eta_T - \eta_R \leq \epsilon_L$$

$$H_{a2}: \quad \eta_T - \eta_R > \epsilon_L.$$

For each pair of hypotheses, a test statistics is computed and then compared to a critical value from the t distribution. Specifically, H_{01} is rejected if

$$\frac{\bar{y}_T - \bar{y}_R - \epsilon_U}{s\sqrt{\frac{1}{n_T} + \frac{1}{n_R}}} < -t_{\alpha,r},$$

and H_{02} is rejected if

$$\frac{\bar{y}_T - \bar{y}_R - \epsilon_L}{s\sqrt{\frac{1}{n_T} + \frac{1}{n_R}}} > t_{\alpha,r},$$

where $\bar{y}_T$ denotes the sample mean of a random sample for the test drug from a normal population with mean η_T and variance σ^2, and $\bar{y}_R$ denotes the sample mean of a random sample for the reference drug from a normal population with mean η_R and variance σ^2, n_T and n_R are the sample sizes for the test and reference group, and s^2 is the pooled estimate of variance σ^2, i.e., $s^2 = \{\sum_{i=1}^{n_T}(y_i - \bar{y}_T)^2 + \sum_{j=1}^{n_R}(y_j - \bar{y}_R)^2\}/(n_T + n_R - 2)$. Here, $t_{\alpha,r}$ is the upper 100α percentile of a t distribution with $r = n_T + n_R - 2$ degrees of freedom.

These equations are highly similar to the familiar Student's t-statistic for traditional, one-sided tests of statistical significance with the only difference being that the lower equivalence bound ϵ_L and the upper equivalence bound ϵ_U are subtracted from the mean difference between groups. If we successfully reject the null for both tests, practical equivalence may then be asserted at the $100\alpha\%$ significance level; otherwise, we fail to reject the assumption of a practical difference.

3.2.2 Confidence Interval Approach to Equivalence Testing

The TOST procedure is also identical to forming the corresponding $(1-2\alpha)100\%$ confidence interval and declaring the two groups equivalent if the interval lies entirely within the tolerance limits (ϵ_L, ϵ_U). The reason the confidence level is $(1-2\alpha)100\%$ and not the usual $(1-\alpha)100\%$ is because this method is equivalent to performing two separate one-sided level-α tests. A careful explanation can be found in Berger and Hsu (1996) and Barker et al. (2001). Thus, a 90% confidence interval yields a 5% significance level for testing equivalence. Both FDA (1992) and EC (1993) specify that the TOST should be executed in this fashion. Equivalence for differences between two independent means shown in the previous section will be declared if the lower confidence bound $\bar{y}_T - \bar{y}_R - t_{\alpha,r} s\sqrt{1/n_T + 1/n_R}$ is above ϵ_L and the upper confidence bound $\bar{y}_T - \bar{y}_R + t_{\alpha,r} s\sqrt{1/n_T + 1/n_R}$ is below ϵ_U.

When both the tests of difference and equivalence tests are used, there will be four possible outcomes in the confidence interval procedure: the effect is statistically equivalent and not statistically different from zero (Scenario A), statistically equivalent and statistically different from zero (Scenario B), statistically different from zero but not statistically equivalent (Scenario C), or neither statistically different from zero nor statistically equivalent (Scenario D). In Figure 3.2, mean differences (black dots) and their 90% (thick lines with arrows) and 95% (thin lines) confidence intervals are illustrated for the four scenarios. The equivalence limits chosen here are $\epsilon_L = -0.3$ and $\epsilon_U = 0.5$.

To conclude equivalence (Scenario A and B), the 90% CI around the observed mean difference should fall within the equivalence bands $(\epsilon_L, \epsilon_U) = (-0.3, 0.5)$. The traditional two-sided test of difference is rejected (Scenario C) when the CI for the mean difference does not include zero. Effects can be statistically different from zero and statistically equivalent (Scenario B) when the 90% CI falls within the equivalence bounds and the 95% CI excludes zero. Finally, an effect can be neither statistically different from zero nor statistically equivalent (Scenario D) when the 90% CI includes one of the equivalence bounds and the 95% CI

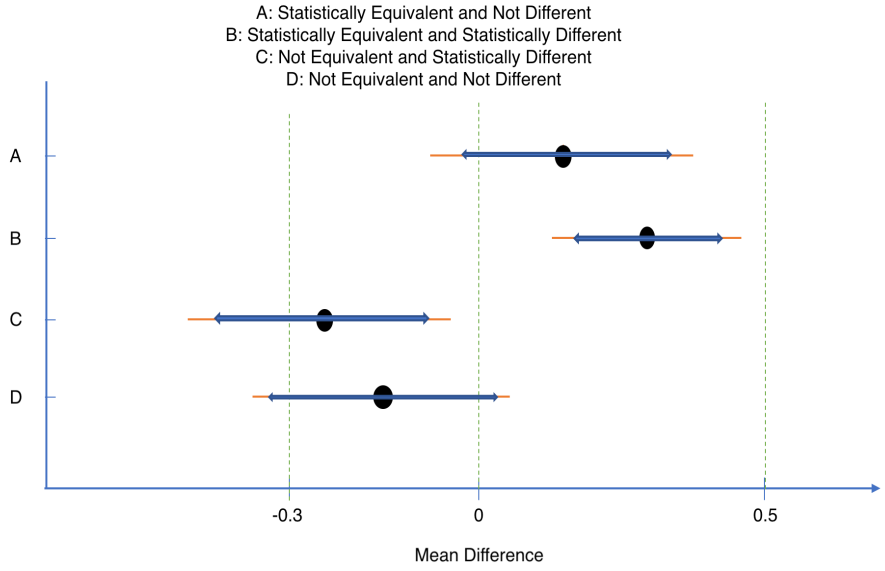

Figure 3.2: *Mean differences (black dots) and 90% confidence intervals (thick blue horizontal lines with arrows) and 95% confidence intervals (thin orange horizontal lines) with equivalence bands* $\epsilon_L = -0.3$ *and* $\epsilon_U = 0.5$ *for four combinations of test results that are statistically equivalent or not and statistically different from zero or not.*

includes zero.

3.2.3 Setting Equivalence Bands

The determination of equivalence limits, ϵ_L and ϵ_U, is the most critical step in equivalence testing. Large values of ϵ_L and ϵ_U determine a wider equivalence region and make it less difficult to establish equivalence. The equivalence limits not only determine the results of the test but also gives scientific credibility to the study. The value of a study or experiment largely depends on how well the equivalence limits can be scientifically justified.

FDA and Center for Drug Evaluation and Research (CDER) recommend that statistical analysis for pharmacokinetic measures, such as AUC and C_{max}, be based on the TOST procedure to determine whether the average values for a pharmacokinetic measure observed after administration of the Test and Reference products are comparable. This approach is

termed average BE and involves the calculation of a 90% confidence interval for the ratio of the averages (population geometric means) of the measures for the Test and Reference drug. To establish average BE, the calculated confidence interval should fall within the BE limits, usually 80%-125% for the ratio of the product averages.

For a broad range of drugs, a BE limit of 80% to 125% for the ratio of the product averages has been adopted for the use of an average BE criterion. Generally, this limit of 80%-125% is based on a clinical judgment that a test product with measures outside this range should be denied market access. However, for highly variable drugs, defined as those for which the within-subject variability equals or exceeds 30% of the mean C_{max} and/or AUC, the equivalence limits may be widened to 0.75 -1.33 or even 0.70-1.43 (Midha et al., 2005). We will discuss the procedure of determining appropriate equivalence bands for our study in Section 4.4.

3.3 Equivalence Testing for Functional Data

Equivalence testing for scalar data has been well-addressed in the literature. However, the same cannot be said for functional data. The complication from keeping the functional structure of the data, rather than using a scalar transformation to reduce dimensionality, makes the existing literature on equivalence testing deficient for the desired inference on pulse waveforms. In this work, a framework for functional equivalence testing that is analogous to its univariate counterpart is proposed and evaluated.

3.3.1 TOST for Functional Data

A TOST procedure is a commonly used approach for conducting the hypothesis testing within the frequentist paradigm. Brown et al. (1997) and Berger and Hsu (1996) propose procedures which are uniformly more powerful for the scalar data since the TOST procedure

can suffer from a lack of power. However, these methods are themselves quite complicated even for univariate data, and their added complexity does not seem to be justified by their relatively small gain the power (Meyners, 2012). For this reason, in the majority of applications, TOST continues to be the method of choice extended to the equivalence testing for functional data within the frequentist paradigm.

In the functional data setting, let $\theta(\cdot)$ denote a functional measurement of similarity between the location parameters of two functions. One potential choice for $\theta(\cdot)$ is the difference between overall mean functions, $\mu_1(\cdot) - \mu_2(\cdot)$. Let $\kappa_l(\cdot)$ and $\kappa_u(\cdot)$ denote lower and upper equivalence limits which are now functions. These bands are chosen such that practical equivalence can be established depending on whether or not $\theta(\cdot)$ falls entirely within the equivalence bands defined by $\kappa_l(\cdot)$ and $\kappa_u(\cdot)$. The null and alternative hypotheses for the tests of location can be stated as follows:

$$H_0^\theta : \exists t \in \mathcal{T}, \theta(t) \notin (\kappa_l(t), \kappa_u(t)),$$
$$H_a^\theta : \forall t \in \mathcal{T}, \theta(t) \in (\kappa_l(t), \kappa_u(t)).$$

Note that the above test, in aggregate, is an Intersection-Union Test (IUT)(Berger and Hsu, 1996). To test these hypotheses within the frequentist paradigm, we can conduct TOST procedures for the location parameter $\theta(\cdot)$. Since this is an IUT, each of the hypothesis tests can be conducted at significance level α to arrive at an overall size of α. In practice, functional data are measured along a finite grid of values. Thus, the grid must be fine enough so that areas of potential dissimilarity along the domain are not ignored.

3.3.2 Nonparametric Bootstrap Functional Equivalence Test

Fogarty and Small (2014) proposed using the nonparametric bootstrap of Efron and Tibshirani (1994) for assessing equivalence by constructing point-wise confidence intervals for the metric of equivalence and then using the duality between confidence intervals and point-wise hypothesis tests to conduct the inference. We modified the nonparametric bootstrap approach and applied it to the simulated data. If we consider two sets of curves generated from the same person, then equivalence should be concluded since these two sets share the same population mean.

We define the first set of curves as the reference group, and the second set of curves as the test group. Let $\mu_R(t)$ denotes the population mean curve for the reference group and $\mu_T(t)$ denote the population mean curve for the test group. Consider the ratio of two mean curves as a metric for equivalence and then conduct the following hypotheses:

$$H_0 : \frac{\mu_T(t)}{\mu_R(t)} \geq \theta_u \quad or \quad \frac{\mu_T(t)}{\mu_R(t)} \leq \theta_l \quad for\ some\ t$$
$$H_a : \theta_l \leq \frac{\mu_T(t)}{\mu_R(t)} \leq \theta_u \quad for\ all\ t$$

The values of θ_l and θ_u are specified by the researcher and define how close the mean curves must be to be declared equivalent. We might choose $\theta_l = 0.8$ and $\theta_u = 1.25$ as the equivalence limits as suggested by the FDA guidance on BE. If we can show that the ratio of two curves falls within the range of (θ_l, θ_u), the underlying mean curves are declared to be equivalent. Let $\hat{y}_R(\cdot)$ and $\hat{y}_T(\cdot)$ denote the nonparametric regression estimates of repeated measurement curves for the reference group and test group, respectively. We use

$$T(\cdot) = \frac{\hat{y}_T(\cdot)}{\hat{y}_R(\cdot)}$$

as test statistics for the hypothesis and apply the nonparametric bootstrap procedure to de-

rive point-wise confidence intervals. The duality between one-sided confidence intervals and one-sided tests helps to decide to reject or fail to reject the difference of the null hypothesis. Let $y_{Ri}(t)$ be the i^{th} curve for the reference group, $i = 1, 2, \cdots, n_R$, and $y_{Ti}(t)$ be the i^{th} curve for the test group, $i = 1, 2, \cdots, n_T$. We compute the sample averages for both the reference and the test group at each time points $\bar{y}_R(t) = \frac{1}{n_R}\sum_{i=1}^{n_R} y_{Ri}(t)$ and $\bar{y}_T(t) = \frac{1}{n_T}\sum_{i=1}^{n_T} y_{Ti}(t)$.

In each iteration, we do the following:

1. Sample n_R curves with replacement from the reference group, and sample n_T curves with replacement from the test group.

2. Compute nonparametric regression estimates of the repeated measurements curves for the reference and test group using the *sm.rm* function from R package **sm** (Bowman and Azzalini, 2015). Denote these as $\{\hat{y}^*_R(\cdot)\}$ and $\{\hat{y}^*_T(\cdot)\}$. We compute $\hat{y}^*_R(\cdot) = S^*_R\bar{y}^*_R(\cdot)$ and $\hat{y}^*_T(\cdot) = S^*_T\bar{y}^*_T(\cdot)$, where $\bar{y}^*_R(\cdot)$ and $\bar{y}^*_T(\cdot)$ are the sample average curves, and S^*_R and S^*_T are the smoothing matrices depending on the smoothing parameter for the reference and the test group, respectively. Hart and Wehrly (1986) proposed a criterion for the choice of the smoothing parameter in the context of repeated measurements.

3. Compute $\hat{T}^*(\cdot) = \frac{\hat{y}^*_T(\cdot)}{\hat{y}^*_R(\cdot)}$.

4. Store this value.

We perform N iterations as shown above and record the bootstrap test statistics $\hat{T}^*(\cdot)$. The upper and lower point-wise confidence intervals for $T(t)$ using a bias correcting percentile-based bootstrap as discussed in Davison and Hinkley (1997) is:

$$C^u_{1-\alpha}(T(t)) = [2\hat{T}(t) - q_\alpha[\hat{T}^*(t)], \infty),$$
$$C^l_{1-\alpha}(T(t)) = (-\infty, 2\hat{T}(t) - q_{1-\alpha}[\hat{T}^*(t)]),$$

where $\hat{T}(t)$ is the observed test statistic and $q_\alpha[\hat{T}^*(t)]$ is the αth percentile of $\hat{T}^*(t)$ we computed from step 3 in each iteration. If our lower equivalence limit at time t, θ_l, is

outside of $C^u_{1-\alpha}(T(t))$, then we can reject the null that $T(t) < \theta_l(t)$ at point t. Likewise, if $\theta_u(t)$, is outside of $C^l_{1-\alpha}(T(t))$, we reject the null of $T(t) > \theta_u(t)$ at point t. The final equivalence can be declared if we successfully reject both tests at every time point t. We call this nonparametric bootstrap procedure the nonparametric bootstrap functional equivalence test (NBFET) and apply it to the simulated data in Chapter 4 to verify its validity and efficiency.

Chapter 4

Development of Equivalence Bands and Method Validation

Having proposed the NBFET procedure in the previous chapter, we now proceed to evaluate the performance of the proposed testing procedure on the simulated data. We also use our simulations as a way to determine the equivalence bands that vary based on the coefficient of variation (CV). We present the details of our evaluation based on equivalence testing, including the choice of the data generating process, the determination of equivalence bands, the evaluation methodology and the simulation results.

This chapter is organized as follows, in Section 4.1 we give the rationale for using simulation to carry out our evaluation, followed by the justification of the selection of data generating process. We then apply NBFET discussed in Chapter 3 to the simulated data based on common choices of equivalence bands in the drug approval processes in Section 4.3. In Section 4.4, we determine the equivalence bands appropriate for our study from the simulation results. Then, we apply NBFET procedure to simulated data based on the chosen equivalence bands, and we conclude with a discussion in Section 4.5. In Section 4.6, we perform equivalence testing on a small set of feature points to see whether the procedure can be simplified but also reach the same conclusion.

4.1 Purpose of Simulation

For our project, we have appropriate and well-behaved data from only four persons. Moreover, there is a lot of variability and inaccuracy at the beginning of the curves due to the difficulty of separating each pulse wave cycles. By observing simulated outcomes, we can gain insight into the behavior of real-life pulse waves without subjecting a large number of volunteers to repeated collection of their pulse waves. If the methodology works well for the simulated data, it is more likely to run for the real-life situation. The following are some advantages simulation study will bring to our project.

• Simulation provides complete control over the environment. For example, during evaluation, we like to set the curves to start exactly at zero with increased variability over time. This is easy to do with simulation.

• Simulation allows sharing the code with others. Thus, the results can be duplicated and verified by other researchers, who may want to evaluate their own scenarios, leading to a more complete evaluation.

• With simulation, other types of functional data can also be generated using the same methodology and procedure. Hence, it allows for setting up a better and more meaningful evaluation schema.

4.2 Generating Simulated Functional Data using Brownian Motion

In this section, we introduce how to generate the pulse wave behavior-like curves from the Brownian motion (Hida, 1980) for use in the simulation. Brownian motion was proposed to describe the random movement of particles in a fluid due to their collisions with other fast-moving molecules in the fluid. Brownian motion is also known as pedesis, which comes from the Greek word for "leaping". Most examples of Brownian motion are transport processes that are also affected by larger currents, such as the motion of pollen grain on still water, movement of dust motes in a room, diffusion of calcium through bones.

Definition 1. *A stochastic process* $\{X(t), t \geq 0\}$ *is said to be a Brownian motion process if*
(i) $X(0)=0$*;*
(ii) $\{X(t), t \geq 0\}$ *has stationary and independent increments;*
(iii) for every $t > 0$*,* $X(t)$ *is normally distributed with mean 0 and variance* $\sigma^2 t$*.*

From the definition of the Brownian motion, we can see that the process always starts at zero, and for every time point t, it follows a normal distribution with mean zero and a variance which is a linear function of t. Thus, there is smaller variability at the beginning of the curves and more considerable variability at the tail. Those are features we observe from PPG pulse waves. Therefore, we propose to simulate curves that mimic the behavior of pulse waves from Brownian motion with a known underlying population mean. A mathematical model that describes Brownian motion is the Wiener process, a continuous-time stochastic process named in honor of Norbert Wiener. It has been widely used in math, economics, engineering, physics, biology, chemistry, etc.

Definition 2. *A stochastic process* $\{X(t), t \geq 0\}$ *is said to be a Brownian motion process with drift coefficient* μ *and variance parameter* σ^2 *if*

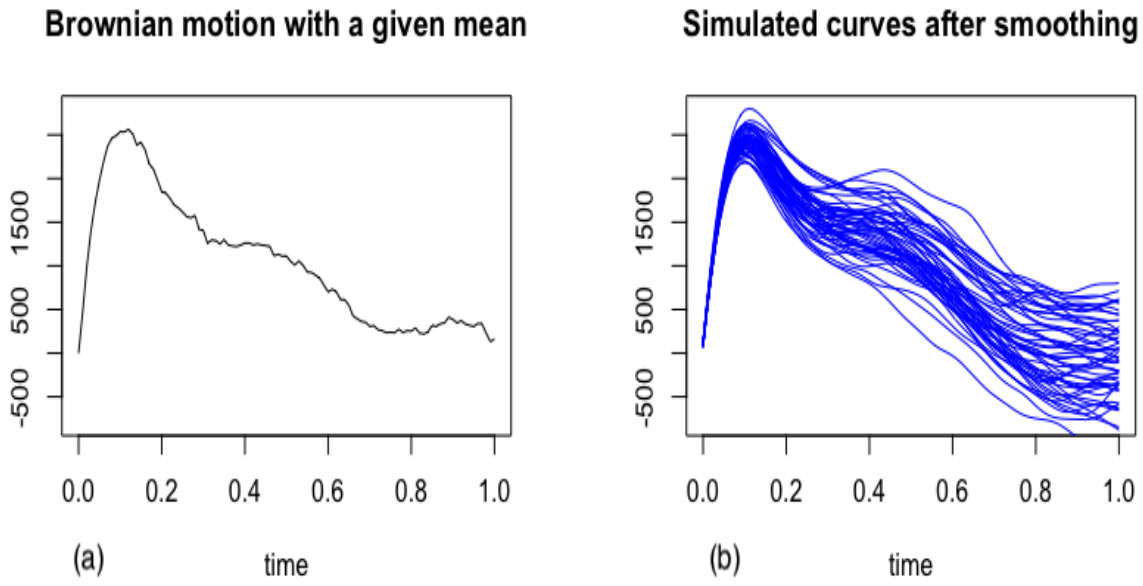

Figure 4.1: *(a) A simulated curve from the Brownian motion with mean* $500f(t; 2.1, 12) + 550f(t; 4, 5)$ *and variance 40 , (b) a group of 50 simulated curves from Brownian motion after smoothing using nonparametric regression.*

(i) X(0)=0;

(ii) $\{X(t), t \geq 0\}$ *has stationary and independent increments;*

(iii) for every $t > 0$*, X(t) is normally distributed with mean* μt *and variance* $\sigma^2 t$*.*

We generate the curves from a modified version of *Brownian motion with drift*, where our mean is not linear in t but a function of t. Let $f(t; \alpha, \beta)$ denote the probability density function of a Beta distribution with shape parameters α and β, where $0 \leq t \leq 1$ and $\alpha, \beta > 0$. Figure 4.1 (a) shows a simulated curve from the Brownian motion with mean $500f(t; 2.1, 12) + 550f(t; 4, 5)$ and variance $\sigma^2 = 40$, and (b) shows a realization of 50 simulated curves after smoothing using nonparametric regression. From Figure 4.1 (b), we can see that the simulated curves mimic the behavior of the pulse waves and have similar patterns we found in real-life data after preprocessing.

4.3 Nonparametric Bootstrap with Common Choices of Equivalence Bands

We consider the curves generated from Brownian motion with the following means as the simulated data :

Mean A: $500f(t; 2.1, 12) + 550f(t; 4, 5)$,

Mean B: $500f(t; 2.3, 12) + 460f(t; 4.04, 5.1)$,

Mean C: $500f(t; 2.3, 12) + 500f(t; 4, 5)$,

Mean D: $250f(t; 2.8, 12) + 520f(t; 2, 2)$,

Mean E: $270f(t; 2.7, 12) + 490f(t; 2.1, 2)$,

Mean F: $500f(t; 2.1, 12) + 550f(t; 3, 3)$,

where $f(t; \alpha, \beta)$ is defined in Section 4.2. Namely, the underlying population means are a mixture of beta distributions. The beta probability density function is unimodal when both shape parameters are greater than one, and we can therefore control where the mode occurs by changing the shape parameters. However, there are two peaks in the real pulse waves. Therefore, we simulated curves from Brownian motion with population means that are a mixture of two beta densities where each distribution has both shape parameters greater than one.

Figure 4.2 shows the six population means used in our simulation study. These mean curves represent some of the situations we will encounter in real life. For example, Mean A and Mean C as are Mean D and Mean E are quite close to each other the entire time range; while Mean A and Mean F behave similarly near the first peak but are disparate after the notch. Mean D and Mean E are quite different from the other curves. They also show similar patterns we found in real pulse waves. We will apply NBFET as discussed in Section 3.3.2 to the simulated curves from Brownian motion with means shown above to measure the equivalence.

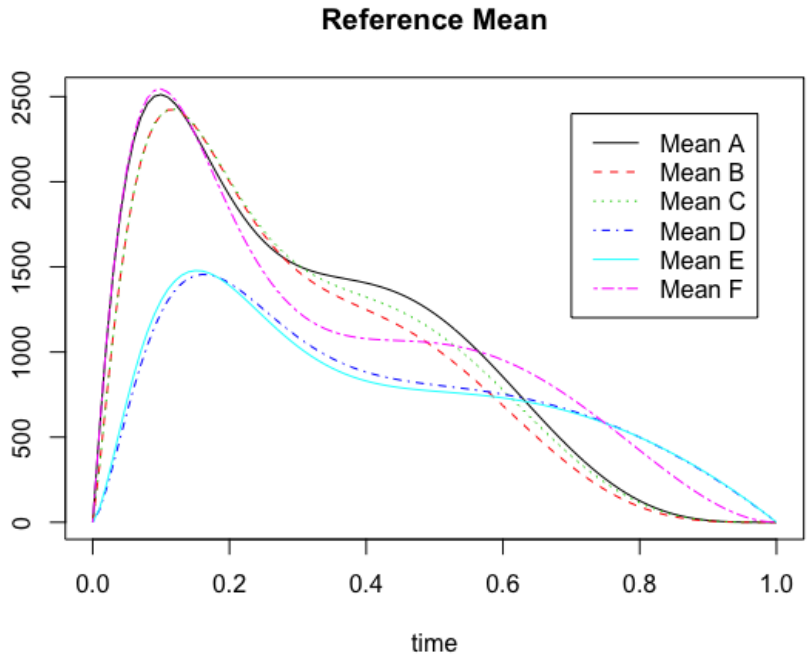

Figure 4.2: *The underlying population means for the simulation study.*

4.3.1 Equivalence Bands

Recall the assessment of bioequivalence discussed in Section 3.2, the two drugs are bioequivalent if the population means of pharmacokinetic characteristics AUC, C_{max}, or t_{max} are sufficiently close. For example, let μ_T denote the population mean AUC for the generic (Test) drug and μ_R denote the population mean AUC for the name brand (Reference) drug. To demonstrate bioequivalence, the following hypotheses are tested:

$$H_0: \quad \frac{\mu_T}{\mu_R} \geq \theta_U \ \ or \ \frac{\mu_T}{\mu_R} \leq \theta_L$$
$$H_a: \quad \theta_L < \frac{\mu_T}{\mu_R} < \theta_U.$$

The values of θ_L and θ_U are specified by the researchers and define how "close" the mean responses of the measured characteristic from two drugs must be to be declared bioequivalent. The interval (θ_L, θ_U) is called the zone of indifference or equivalence bands. According

to the FDA, average bioequivalence is demonstrated if the 90% confidence interval of the geometric mean ratios of the test and the reference drugs for the AUC fall completely within limits (in %) of 80-125%. In other words, $\theta_L = 0.8$ and $\theta_U = 1.25$, which are symmetric in the ratio scale.

Even though a review of more than a decade of bioequivalence data from the FDA supports the average bioequivalence criteria in approving high-quality drugs, there are some concerns over some generic drugs that have a narrow therapeutic window such as some anti-epileptic and anti-coagulant drugs. Many authors have criticized the one-fits-all criterion because it does not consider the therapeutic window and variability of a drug. A highly variable reference drug, which is defined as a drug with within-subject coefficient of variation (CV) in one or more of the pharmacokinetic characteristics being 30% or larger, may not be demonstrated to be bioequivalent even to itself in a typical cross-over study with a moderate number of subjects (Davit et al., 2012). A review of 1010 bioequivalence studies of 180 generic drugs submitted to the FDA during 2003-2005 suggests that 31%(57/180) of those are highly variable. Therefore, it is necessary to develop an alternative to the usual (0.8, 1.25) rule for highly variable drugs. It has been suggested by the European Agency for the Evaluation of Medicinal Products (EMEA) that the bioequivalence bands of the 90% confidence interval could be widened from 0.8-1.25 to 0.75-1.33 or even to 0.7-1.43, depending on the within-subject CV. For our project, we will determine the equivalence bands appropriate for different ranges of CV through a simulation study.

4.3.2 Determining the Evaluation Range

Before we determine the equivalence bands that are appropriate for our project through simulation, we truncate the curves on the left and right, and limit our analysis on a specific evaluation range. It would be the best if we could perform the equivalence test on the entire $(0, 1)$ time range; however, it causes some issues. First, each curve starts at zero and quickly

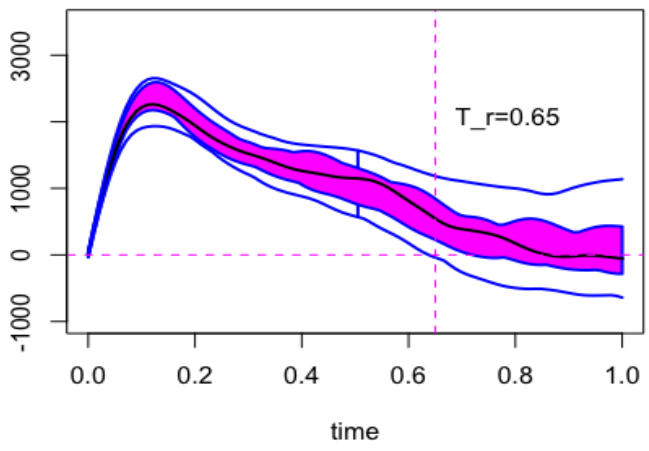

Figure 4.3: *Functional boxplot of a group of 50 curves generated from Brownian motion with Mean 3 and variance of 40. We pick the truncation points at [0, 0.65].*

increases at the beginning part. Thus, ratio of the test mean to the reference mean rapidly goes up toward infinity early on even though the values are extremely close together. Second, there is a lot of variability at the end of the curves, and the CVs at these points are bigger than one. It is not very meaningful to measure the equivalence when the standard deviation is larger than the mean. Third, some curves go below zero at the tail, and we have negative values of the ratio of the mean curves. It is possible to have negative ratios but not practically meaningful for our project. Therefore, we truncate the curves and make the equivalence measurement within a specific range. The majority of meaningful information for authentication is contained between the systolic peak and the dicrotic notch of the pulse waves. Thus, we will make sure that we retain that part of the waveform for testing equivalence.

Let $\hat{y}(t)$, where $t \in (0, 1)$, denote the estimated smoothing curve using nonparametric regression for the reference data. We first determine the evaluation range $[T_1, T_r]$ where the majority curves from the reference group have positive values. The values of T_1 and T_r can be determined from the functional boxplot of the reference data where the maximum non-

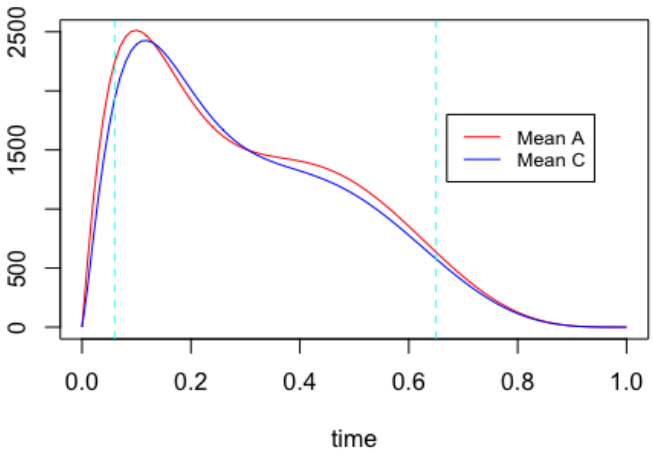

Figure 4.4: *Population Mean A and Mean C with the evaluation range of* $[0.06, 0.65]$.

outlying envelope is almost all positive. Let D_t be the time point where the first derivative of the sample mean curve $\hat{y}(t)$ gets the maximum, and M_t be the point where the mean curve $\hat{y}(t)$ reaches the first peak (maximum). We define the largest value among T_1, D_t, and $\frac{1}{2}M_t$ as the left truncation point T_l, namely, $T_l = max\{T_1, D_t, \frac{1}{2}M_t\}$. We then apply NBFET on the evaluation range $[T_l, T_r]$. For example, we generate a group of 50 curves from Brownian motion with Mean C and variance of 40 as the reference data and the functional boxplot is shown in Figure 4.3. We will select $[T_1, T_r] = [0, 0.65]$ as the initial range since its non-outlying envelope is almost all positive. By take the first derivative of the sample mean curve $\hat{y}(t)$, we get $D_t = 0.05$ and $M_t = 0.12$. Therefore, the left truncation point $T_l = max\{0, 0.05, \frac{1}{2} * 0.12\} = 0.06$, and the evaluation range $[T_l, T_r] = [0.06, 0.65]$. Figure 4.4 shows the population mean curves from Mean A and Mean C, with the evaluation range $[0.06, 0.65]$. Table 4.1 shows an example of the evaluation ranges for each of the reference mean with a specific value of average of CV. For example, if we generate a group of curves from a Brownian motion with Mean A and variance of 40 as the reference data, we will apply the equivalence tests on the evaluation range of $[0.06, 0.6]$.

Reference Mean	Variance of Brownian motion	Average of CV	Evaluation range
Mean A	40	0.146	[0.06, 0.6]
Mean B	40	0.125	[0.075, 0.6]
Mean C	40	0.172	[0.07, 0.65]
Mean D	25	0.162	[0.095, 0.65]
Mean E	25	0.136	[0.095, 0.65]
Mean F	40	0.141	[0.06, 0.6]

Table 4.1: *The evaluation range for each of the reference mean with a specific value of average of CV.*

	Ref Data	Test Data	average of CV	EBs	# rejecting	Clopper-Pearson CI
H_a True	Mean A Var=40	Mean A Var=40	0.146	(0.8, 1.25)	999/1000	(0.994, 0.999)
				(0.75, 1.33)	1000/1000	(0.996, 1.000)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean A Var=80	Mean A Var=80	0.244	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	965/1000	(0.952, 0.976)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean A Var=100	Mean A Var=100	0.314	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	660/1000	(0.629, 0.689)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean A Var=120	Mean A Var=120	0.486	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	26/1000	(0.0171, 0.0379)
				(0.7, 1.43)	992/1000	(0.984, 0.997)

Table 4.2: *The null hypothesis of difference is false since the reference and the test data are generated from the same mean. We compute the powers of the tests for different average of CV and EBs. EB represents equivalence band and CI represents confidence interval.*

4.3.3 Simulation Results for Various Equivalence Bands

Let $\mu_R(t)$ denote the population mean curve for the reference group and $\mu_T(t)$ represent the population mean for the test group. We consider the ratio of two mean curves as a metric for equivalence and conduct the following hypotheses:

$$H_0 : \frac{\mu_T(t)}{\mu_R(t)} > U(t) \ or \ \frac{\mu_T(t)}{\mu_R(t)} < L(t)$$
$$H_a : L(t) \leq \frac{\mu_T(t)}{\mu_R(t)} \leq U(t)$$

where $(L(t), U(t))$ is the predefined equivalence band (EB).

We first consider the case where the reference and the test data are from the same known population mean. In this case, the null hypothesis of difference is false since the reference, and the rest data are actually from the identical mean. For example, we generate groups of fifty curves from a Brownian motion with Mean A as both the reference and the test data. We then apply NBFET to the curves within the evaluation range discussed in Section 4.3.2, and the results are shown in Table 4.2. From the first part of Table 4.2, we can see that when we use a variance of 40 in the Brownian motion, namely, the average of point-wise CVs for the reference data is 0.146, we successfully reject H_0 of difference 999 out of 1000 times when the equivalence band is $(0.8, 1.25)$, and Clopper-Pearson confidence interval for power is (0.994, 0.999). With wider equivalence margins of $(0.75, 1.33)$ or $(0.7, 1.43)$, we reject all 1000 times and get observed power of 100%. We also consider other options of variances, denoted by σ^2, to cover a wide range of CVs for real-life situations. We can see that when the variance of Brownian motion increases from 40 to 80, the average of point-wise CVs goes up to 0.244, the power of the test decreases from 0.999 to 0 if the equivalence band is $(0.8, 1.25)$. With wider equivalence band of $(0.75, 1.33)$, the power of the equivalence test goes down from 100% to 96.5%. When the average of the point-wise CVs gets larger say, 0.486, the narrower equivalence bands of $(0.8, 1.25)$ and $(0.75, 1.33)$ do not provide good power even though the reference and the test data are actually from the same population mean. However, the power is still high (99.2%) when we use the widest bands of $(0.7, 1.43)$. Table 4.3 show the simulation results when the reference and the test data are generated from Brownian motion with Mean B, Mean C, and Mean D. The complete simulation results are shown in Appendix A. From the simulation, the equivalence bands of $(0.8, 1.25)$ works well when the average of CVs is smaller than 0.2, however, when the CVs increase to 0.2 but are lower than 0.3, we may enlarge the equivalence bands to $(0.75, 1.33)$. It will be appropriate to use the equivalence bands of $(0.7, 1.43)$ when the average of CVs is beyond 0.3.

	Ref Data	Test Data	average of CV	EBs	# rejecting	Clopper-Pearson CI
H_a True	Mean B Var=40	Mean B Var=40	0.125	(0.8, 1.25)	998/1000	(0.993, 0.999)
				(0.75, 1.33)	1000/1000	(0.996, 1.000)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean B Var=60	Mean B Var=60	0.218	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	935/1000	(0.918, 0.949)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean B Var=100	Mean B Var=100	0.334	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	522/1000	(0.491, 0.553)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean B Var=120	Mean B Var=100	0.467	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	2/1000	(0.006, 0.021)
				(0.7, 1.43)	990/1000	(0.981, 0.995)
H_a True	Mean C Var=40	Mean C Var=50	0.172	(0.8, 1.25)	997/1000	(0.991, 0.999)
				(0.75, 1.33)	1000/1000	(0.996, 1.000)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean C Var=80	Mean C Var=80	0.225	(0.8, 1.25)	46/1000	(0.034, 0.061)
				(0.75, 1.33)	972/1000	(0.960, 0.981)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean C Var=100	Mean C Var=100	0.313	(0.8, 1.25)	4/1000	(0.00109, 0.0102)
				(0.75, 1.33)	489/1000	(0.458, 0.520)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean C Var=120	Mean C Var=120	0.429	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	147/1000	(0.126, 0.170)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean D Var=25	Mean D Var=25	0.142	(0.8, 1.25)	999/1000	(0.994, 0.999)
				(0.75, 1.33)	1000/1000	(0.996, 1.000)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean D Var=40	Mean D Var=40	0.238	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	892/1000	(0.871, 0.910)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean D Var=50	Mean D Var=50	0.349	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	420/1000	(0.389, 0.451)
				(0.7, 1.43)	944/1000	(0.928, 0.957)
H_a True	Mean D Var=60	Mean D Var=60	0.430	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	34/1000	(0.0237, 0.0472)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)

Table 4.3: *Continued: The null hypothesis of difference is false since the reference and the test data are generated from the same mean. We compute the powers of the tests for different EBs. EB represents equivalence band and CI represents confidence interval.*

4.4 Determination of the Equivalence Bands

Based on the simulation results we got from Section 4.3.3, we decide to use equivalence bands of $(0.8, 1.25)$ when the CV for the reference data at time t, denoted $CV(t)$, is less than 0.2, equivalence bands of $(0.75, 1.33)$ when the CV at time t is between 0.2 and 0.3, and equivalence bands of $(0.7, 1.43)$ when the CV at time t is between 0.3 and 1. It does not make sense to evaluate equivalence when the standard deviation is larger than the mean; thus, we truncate the curves at t when $CV(t) > 1$ for reference data. If $CV(t) > 1$ occurs before the dicrotic notch, we still truncate at t when $CV(t) > 1$ because we believe that the equivalence can hardly be measured when the standard deviation is bigger than the mean.

We define the EB $= (L(t), U(t))$ which determines whether the null hypothesis of difference is true or false. If the ratio of two population means falls entirely within the EB, then the null hypothesis of difference is false. The interval $(L(t), U(t))$ is determined by the CV of the reference data. In other words, for the following hypotheses in an equivalence test,

$$H_0 : \frac{\mu_T(t)}{\mu_R(t)} > U(t) \ or \ \frac{\mu_T(t)}{\mu_R(t)} < L(t)$$
$$H_a : L(t) \leq \frac{\mu_T(t)}{\mu_R(t)} \leq U(t)$$

where $\mu_R(t)$ and $\mu_T(t)$ represent the population mean for the reference and test, respectively. The EB $= (L(t), U(t)) = (0.8, 1.25)$ when $CV(t) < 0.2$, $(L(t), U(t)) = (0.75, 1.33)$ when $0.2 \leq CV(t) < 0.3$, and $(L(t), U(t)) = (0.7, 1.4)$ when $0.3 \leq CV(t) < 1$, where $CV(t)$ is the CV for the reference data at time t. In another words, the equivalence bands are

$$(L(t), U(t)) = \begin{cases} (0.8, 1.25), & \text{if } CV(t) < 0.2 \\ (0.75, 1.33), & \text{if } 0.2 \leq CV(t) < 0.3 \\ (0.7, 1.4), & \text{if } 0.3 \leq CV(t) < 1. \end{cases}$$

The equivalence bands proposed above constitute one of the main contributions in this work. Currently, FDA suggests $(0.8, 1.25)$, which is symmetric in the ratio scale, as the equivalence limits to assess the bioequivalence for a scalar measure of similarity, such as AUC and the maximum concentration. For highly variable drugs, the limits could be widened to $(0.75, 1.33)$ or even to $(0.7, 1.43)$, depending on the within-subject CV. However, these is no literature or guidance in determining the equivalence limits for a functional measurement of similarity.

4.5 Nonparametric Bootstrap with Corrected Equivalence Bands

In this section, we will implement NBFET procedure using the EB determined from Section 4.4. Table 4.4 shows examples where the null hypothesis of difference is false, i.e., the reference and the test data are generated from the population means which are practically equivalent. In the first case, we generate the data from Brownian motion with the same Mean A, and variances of 40, 80, 100, and 120, respectively. The power of the test decreases from 0.999 to 0.89 when the variation increases from 40 to 120 (the average of CV goes up from 0.146 to 0.486). In the second case, the reference and the test data are generated from two different population means Mean E and Mean D, but the null of difference is still false since the ratio of the population means falls entirely within the EB (Figure 4.5). The power of the test goes down from 0.999 to 0.932 when the average of CV increases from 0.162 to 0.416. In the third and fourth cases, the null hypothesis of difference is still false since the ratio of population means is completely contained in the EB. The powers of the tests decrease slightly when the mean values of CV increase. In these four cases, the corrected EB performs well in the sense that the power is maintained at a relatively high level.

Next, consider the case when the null hypothesis of difference is true. Figure 4.6 (a) shows the population Mean A and Mean B, and (b) shows the ratio of Mean A to Mean B within

	Ref Data	Test Data	average of CV	# rejecting	Clopper-Pearson CI
H_a True	Mean A Var=40	Mean A Var=40	0.146	999/1000	(0.9944, 0.9999)
H_a True	Mean A Var=80	Mean A Var=80	0.254	996/1000	(0.9898, 0.9989)
H_a True	Mean A Var=100	Mean A Var=100	0.314	966/1000	(0.9528, 0.9763)
H_a True	Mean A Var=120	Mean A Var=120	0.486	890/1000	(0.8689, 0.9087)
H_a True	Mean E Var=25	Mean D Var=25	0.162	999/1000	(0.9944, 0.9999)
H_a True	Mean E Var=40	Mean D Var=40	0.245	993/1000	(0.9856, 0.9972)
H_a True	Mean E Var=50	Mean D Var=50	0.328	980/1000	(0.9693, 0.9877)
H_a True	Mean E Var=60	Mean D Var=60	0.416	932/1000	(0.9146, 0.9468)
H_a True	Mean A Var=40	Mean C Var=40	0.146	982/1000	(0.972, 0.989)
H_a True	Mean A Var=80	Mean C Var=80	0.254	854/1000	(0.831, 0.875)
H_a True	Mean A Var=100	Mean C Var=100	0.314	945/1000	(0.929, 0.958)
H_a True	Mean A Var=120	Mean C Var=120	0.486	819/1000	(0.794, 0.842)
H_a True	Mean B Var=40	Mean C Var=40	0.125	964/1000	(0.951, 0.975)
H_a True	Mean B Var=80	Mean C Var=80	0.218	938/1000	(0.921, 0.952)
H_a True	Mean B Var=100	Mean C Var=100	0.334	887/1000	(0.866, 0.906)
H_a True	Mean B Var=120	Mean C Var=120	0.467	827/1000	(0.802, 0.850)

Table 4.4: *The power of the equivalence test when the null hypothesis of difference is false, i.e., the population means from the reference and the test group are considered equivalent.*

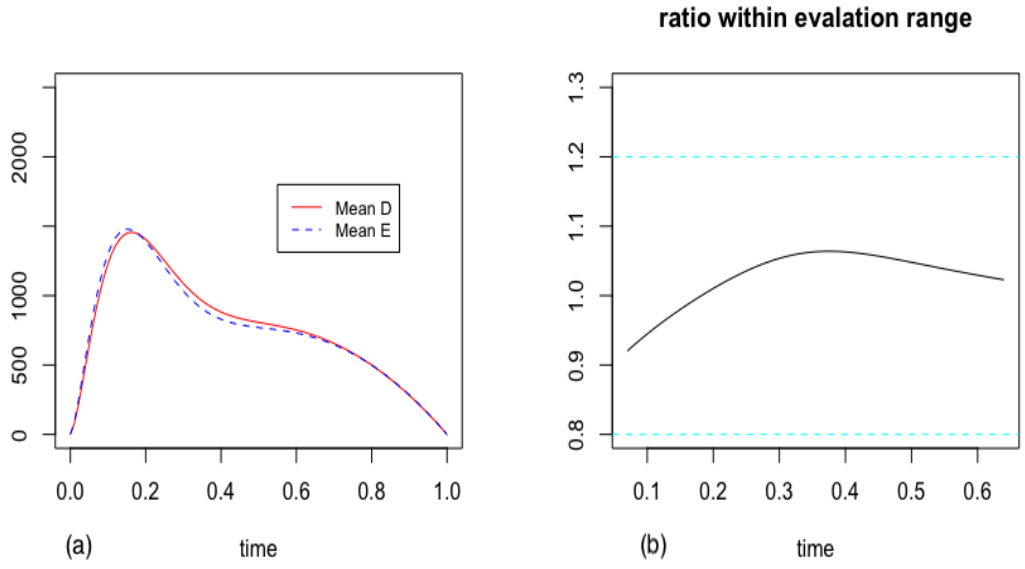

Figure 4.5: *(a) Population Mean D and Mean E. (b) The ratio of Mean D to Mean E within the evaluation range.*

the evaluation range. From Figure 4.6 (b), we can tell that the null hypothesis of difference is true since the ratio of two population means falls just above the upper initial equivalence margin at the tail, and we expect the type I error rate around the significance level of 5%. We compute the type I error rates for different averages of CV and summarize the results in the first part of Table 4.5. When we generate the reference and the test data from Brownian motion with the same variance of 40 (the average of CV is 0.132), we incorrectly reject the null hypothesis of difference 60 out of 1000 times. The type I error rate is 0.06 with the Clopper-Pearson confidence interval $(0.0461, 0.0766)$. We then increase the variance of Brownian motion to 80, 100, and 120, and get the corresponding type I error rates 0.025, 0.046, and 0.019. The type I error rates for all four cases with different averages of CV are controlled around the significance level of 5%. In the second part of Table 4.5, we apply the NBFET to curves generated from Brownian motion with the reference Mean A and the test Mean F. From Figure 4.7, we can see that the null hypothesis of difference is true since the ratio of two means does not fall entirely within the EB. The type I error rate of the test goes down from 0.039 to 0.002 when the mean value of CV increases from 0.146 to 0.486. Figure

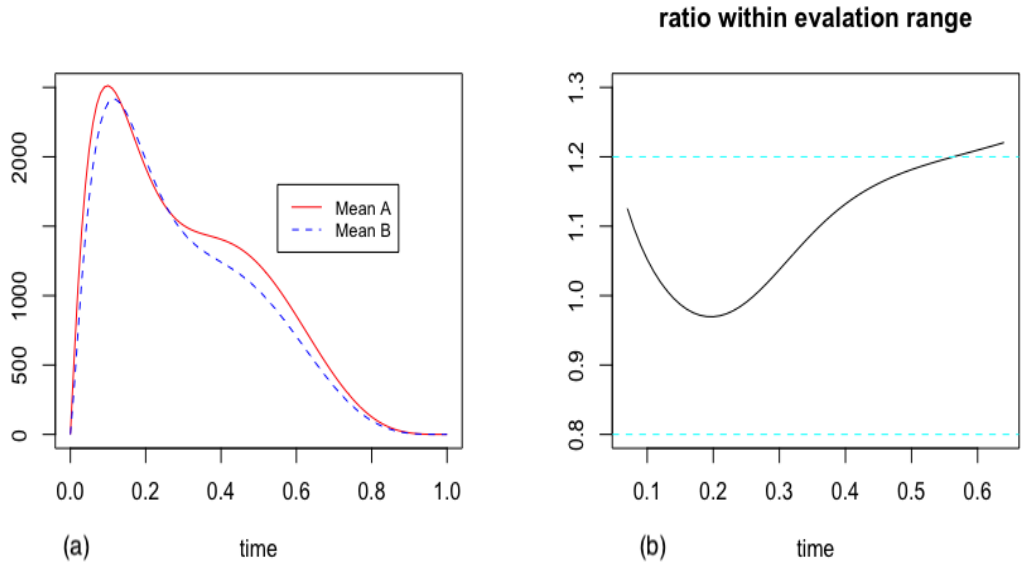

Figure 4.6: *(a) Population Mean A and Mean B. (b) The ratio of Mean A to Mean B within the evaluation range.*

4.8 (a) shows the population Mean B and Mean F, and (b) is the ratio of Mean F to Mean B within the evaluation range. We expect the type I error rate much smaller than the level of significance 5% since the ratio of two means shown in Figure 4.8 (b) is way more extensive than the upper band at the tail. The last part of Table 4.5 confirms this low type I error rate.

The complete simulation results for all situations are shown in Appendix B. From the simulation results based on the corrected EBs, we can see that when the null hypothesis of difference is false, the powers of the test are all above 80%. When the null hypothesis of difference is true, but the ratio of the population means just outside the equivalence boundaries, the type I error rates are controlled around the significance level of 5%. While the type I errors are extremely small if the observed test statistic is way beyond the equivalence boundaries. Therefore, the corrected EBs defined in Section 4.4 work well for the simulated data, and they are likely to work for the real pulse waves as well.

	Ref Data	Test Data	average of CV	# rejecting	Clopper-Pearson CI
H_0 True	Mean B Var=40	Mean A Var=40	0.132	60/1000	(0.0461, 0.0766)
H_0 True	Mean B Var=80	Mean A Var=80	0.268	25/1000	(0.0162, 0.0367)
H_0 True	Mean B Var=100	Mean A Var=100	0.308	46/1000	(0.0339, 0.0609)
H_0 True	Mean B Var=120	Mean A Var=120	0.452	19/1000	(0.0115, 0.0295)
H_0 True	Mean A Var=40	Mean F Var=40	0.146	39/1000	(0.0279, 0.0529)
H_0 True	Mean A Var=80	Mean F Var=80	0.254	21/1000	(0.0130, 0.0319)
H_0 True	Mean A Var=100	Mean F Var=100	0.314	8/1000	(0.00346, 0.0157)
H_0 True	Mean A Var=120	Mean F Var=120	0.486	2/1000	(0.000242, 0.00721)
H_0 True	Mean B Var=40	Mean F Var=40	0.132	3/1000	(0.000619, 0.00874)
H_0 True	Mean B Var=80	Mean F Var=80	0.268	0/1000	(0.00000, 0.00368)
H_0 True	Mean B Var=100	Mean F Var=100	0.308	1/1000	(0.0000253, 0.00556)
H_0 True	Mean B Var=120	Mean F Var=120	0.452	0/1000	(0.00000, 0.00368)

Table 4.5: *The type I error rate of the equivalence test when the null hypothesis is true. The reference and the test data are generated from a Brownian motion with Mean B and Mean A in the first case.*

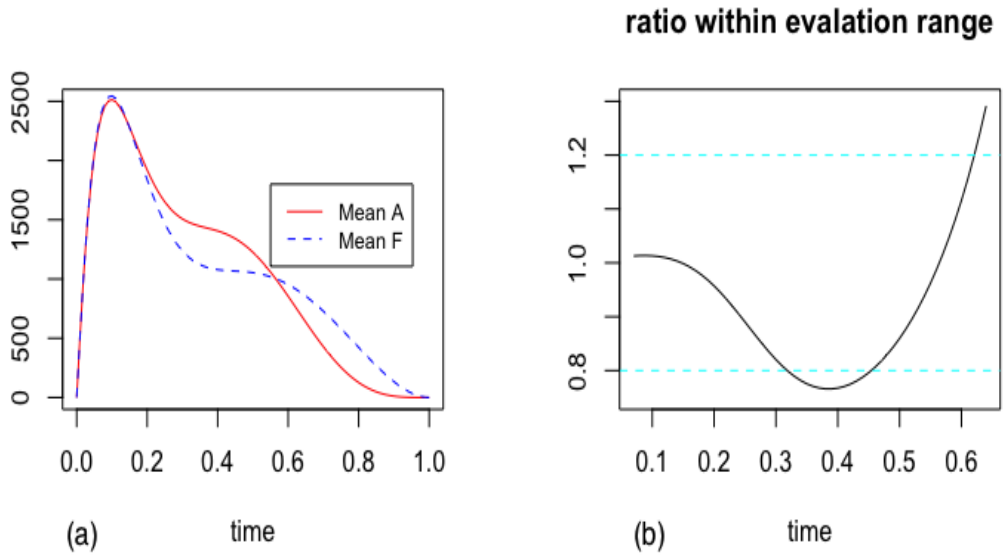

Figure 4.7: *(a) Population Mean A and Mean F. (b) The ratio of Mean F to Mean A within the evaluation range.*

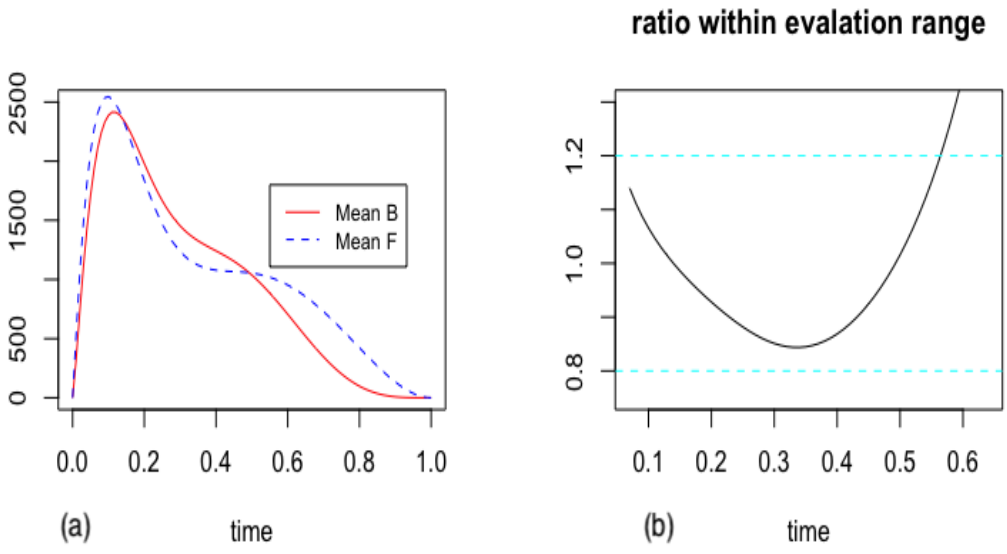

Figure 4.8: *(a) Population Mean B and Mean F. (b) The ratio of Mean F to Mean B within the evaluation range.*

4.6 Simplification by Choosing Featured Points

In the previous sections, we considered the curves as functional data and evaluated the equivalence on the entire time range. We now pick the feature points such as systolic peak, the dicrotic notch, and diastolic peak, and perform the equivalence test on these individual time points. If the equivalence can be concluded for each of the evaluation points, the overall equivalence can then be reached for the whole curve. One of the primary purposes of this section is to see if we can simplify the NBFET procedure we discussed in Section 3.3.2 but also reach the same conclusion by considering only a group of featured evaluation points.

4.6.1 Determining the Featured Evaluation Points

In this section, we will discuss a strategy for selecting the evaluation points for simulated data. Suppose we have a group of p curves as the reference data with known mean and variance of Brownian motion. Let $\hat{y}(t)$, $t \in (0, 1)$ be the nonparametric regression estimate for the reference data. We will perform the following steps to pick the feature points:

Step 1: We determine the evaluation range $[T_l, T_r]$ as discussed in Section 4.3.2.

Step 2: Let M_t be the first peak of the mean curve $\hat{y}(t)$ where its first derivative equals zero. Pick $M_t \pm K_m * \Delta t$ as the first set of evaluation points, where Δt is a very short time interval.

Step 3: Find the changing point (dicrotic notch) of the mean curve $\hat{y}(t)$ where the first derivative goes from increasing to decreasing, and denote it as N_t. Pick $N_t \pm K_n * \Delta t$ as the second set of evaluation points.

We would consider the following combinations of the evaluation points:

Case 1: Few evaluation points: $K_m = K_n = \{0\}$. 2 evaluation points.

Case 2: Moderate evaluation points: $K_m = \{0, 1, 2\}$, $K_n = \{0, 1, 2\}$. 10 evaluation points.

Case 3: Moderate evaluation points: $K_m = \{0, 1, 2\}$, $K_n = \{0, 1, 2, 3, 4, 5\}$. 16 evaluation

points.

Case 4: More evaluation points: $K_m = \{0, 1, 2, 3, 4, 5\}$, $K_n = \{0, 1, 2, \cdots, 9, 10\}$. 32 evaluation points.

We consider only the first peak and dicrotic notch as the evaluation points in Case 1, and extend to its neighborhood in Case 2, 3, and 4. For example, we select a distance of two evaluation points as a neighborhood of both the first peak and the dicrotic notch in Case 2, which results in 10 total evaluation points. Case 1 has the fewest evaluation points and Case 4 has the most. If the evaluation points are outside the range of $[T_l, T_r]$, we use T_l as the left ending point and T_r as the right ending point.

4.6.2 Simulation Results from Simplified Procedure

When the null hypothesis of difference is false and the reference data behaves equivalently as the test data, the Nonparametric Bootstrap with corrected equivalence bands confirmed the equivalence for all within-person and between-people comparisons in Section 4.5. If we consider a smaller set of evaluation points, we will reach to the same conclusion but with a slightly higher power. Therefore, we focus on the case when the null of difference is true and see if evaluation based on a smaller set of measured points will lead us to an incorrect decision.

In the first case of Table 4.5, we considered the situation where the null of difference was true and measured the equivalence between reference data from Mean A and test data from Mean B. For example, if we generated the data from Brownian motion with a variance of 80 for both the reference and test data, the type I error of the equivalence test was 0.025. The type I error increased to 0.046 when the variance of Brownian motion wnet up to 100. Nonparametric Bootstrap with corrected equivalence bands worked well since we controlled the type I error around the significance level of 5%.

Next, we want to see if the procedure could be simplified by measuring the equivalence

	Ref Data	Test Data	average of CV	# rejecting	Clopper-Pearson CI
H_0 True	Mean B Var=40	Mean A Var=40	0.132	Case 1: 1000/1000 Case 2: 1000/1000 Case 3: 986/1000 Case 4: 56/1000	(0.996, 1.000) (0.996, 1.000) (0.977, 0.992) (0.0426, 0.0721)
H_0 True	Mean B Var=80	Mean A Var=80	0.268	Case 1: 998/1000 Case 2: 998/1000 Case 3: 966/1000 Case 4: 33/1000	(0.993, 0.999) (0.993, 0.999) (0.953, 0.976) (0.0228, 0.0460)
H_0 True	Mean B Var=100	Mean A Var=100	0.308	Case 1: 1000/1000 Case 2: 1000/1000 Case 3: 992/1000 Case 4: 41/1000	(0.996, 1.000) (0.996, 1.000) (0.984, 0.997) (0.0296, 0.0552)
H_0 True	Mean B Var=120	Mean A Var=120	0.452	Case 1: 999/1000 Case 2: 999/1000 Case 3: 989/1000 Case 4: 17/1000	(0.994, 1.000) (0.994, 1.000) (0.980, 0.994) (0.00993, 0.0271)

Table 4.6: *The null hypothesis of difference is true, but the simplified procedure concludes the equivalence.*

on a smaller set of evaluation points but still control the type I error. We considered the same data as we used in the first case of Table 4.5 and computed the type I error rates for each of the four cases mentioned in Section 4.6.1. We picked $\Delta t = 0.01$, and the results were shown in Table 4.6. The first part of Table 4.6 showed the situation where the reference and the test data were generated from Brownian motion with Mean A and Mean B, and a common variance of 40 for each of the four cases. Case 1 considered only two evaluation points (the first peak and the dicrotic notch) and got a type I error rate of $1000/1000 = 1$. Case 4 measured equivalence on neighborhoods of the first peak and the dicrotic notch of total 32 points, and controlled the type I error rate at $56/1000 = 0.056$. When we increased the variance of Brownian motion to 80, 100, and 120, only Case 4 maintained the type I error at the desired level. In another word, the procedure failed to distinguish two sets of curves which came from population means that behaved quite differently and incorrectly identified them as equivalent when we reduced the number of evaluation points. We also applied the similar procedure to comparing the curves generated from Brownian motion with the reference Mean A and the test Mean F, and with the reference Mean B and the test

Mean F. The simulation results are shown in Table 4.7. From Table 4.7, we can see that when we increased the number of the featured points from Case 1 to Case 4, only Case 3 and Case 4 gave the type I error rates at the desired level. The procedure could not differentiate two differently behaved means when we decreased the number of the feature points.

The false acceptance of equivalence is sometimes dangerous, for example, it will approve unauthorized access to company's critical business data, or allow an acquaintance to view personal medical records or bank account details. Moreover, the cost of identifying two different subjects as equivalent is much more than incorrectly concluding two similar persons differ. Therefore, we prefer applying equivalent testing on the whole function instead of on a small group of featured evaluation points. However, Case 4 with the most measurement points could be a replacement when the procedure takes too long to execute.

	Ref Data	Test Data	average of CV	# rejecting	Clopper-Pearson CI
H_0 True	Mean A Var=40	Mean F Var=40	0.146	Case 1: 999/1000	(0.994, 0.999)
				Case 2: 785/1000	(0.758, 0.810)
				Case 3: 47/1000	(0.0347, 0.0620)
				Case 4: 17/1000	(0.00993, 0.0271)
H_0 True	Mean A Var=80	Mean F Var=80	0.254	Case 1: 998/1000	(0.993, 0.999)
				Case 2: 564/1000	(0.532, 0.595)
				Case 3: 36/1000	(0.0253, 0.0495)
				Case 4: 21/1000	(0.0130, 0.0319)
H_0 True	Mean A Var=100	Mean F Var=100	0.314	Case 1: 982/1000	(0.972, 0.989)
				Case 2: 611/1000	(0.580, 0.641)
				Case 3: 28/1000	(0.0187, 0.0402)
				Case 4: 2/1000	(0.0000242, 0.00721)
H_0 True	Mean A Var=120	Mean F Var=120	0.486	Case 1: 967/1000	(0.954, 0.977)
				Case 2: 145/1000	(0.124, 0.168)
				Case 3: 4/1000	(0.00109, 0.0102)
				Case 4: 0/1000	(0.00000, 0.00368)
H_0 True	Mean B Var=40	Mean F Var=40	0.132	Case 1: 999/1000	(0.994, 0.999)
				Case 2: 983/1000	(0.973, 0.990)
				Case 3: 117/1000	(0.0977, 0.139)
				Case 4: 3/1000	(0.000619, 0.00874)
H_0 True	Mean B Var=80	Mean F Var=80	0.268	Case 1: 999/1000	(0.994, 0.999)
				Case 2: 864/1000	(0.841, 0.885)
				Case 3: 69/1000	(0.0541, 0.0865)
				Case 4: 1/1000	(0.0000253, 0.00556)
H_0 True	Mean B Var=100	Mean F Var=100	0.308	Case 1: 992/1000	(0.984, 0.997)
				Case 2: 851/1000	(0.827, 0.873)
				Case 3: 2/1000	(0.0000242, 0.00721)
				Case 4: 0/1000	(0.00000, 0.00368)
H_0 True	Mean B Var=120	Mean F Var=120	0.452	Case 1: 989/1000	(0.980, 0.994)
				Case 2: 325/1000	(0.296, 0.355)
				Case 3: 0/1000	(0.00000, 0.00368)
				Case 4: 0/1000	(0.00000, 0.00368)

Table 4.7: *Continued: The null hypothesis is true, but the simplified procedure concludes the equivalence.*

Chapter 5

Conclusion

In the previous chapter, we introduced a procedure to determine the proper equivalence bands from a simulation study and then applied the nonparametric bootstrap to curves simulated from Brownian motion. The type I error rates of the test were controlled around the nominal significance level of 5% and the powers were all above 80%. Here we apply the nonparametric bootstrap procedure discussed in Section 3.3.2 to real pulse waves from four of the persons to determine whether equivalence tests can actually perform identity authentication. The real data applications and conclusions are summarized in Section 5.1. We give a summary of contributions of this work in Section 5.2 and propose some future work in Section 5.3.

5.1 Real Data Application

Recall the four representative persons we used in Chapter 2: Person 1, Person 2, Person 7, and Person 10. There are 57 pulse cycles from Person 1. Figure 5.1 shows the original curves after truncating and shifting and the corresponding functional boxplot. We divide the curves from Person 1 into two parts: the first 30 curves serve as the reference group and the remaining 27 curves are taken to be the test group. From Table 2.2, we notice that the p-value of the permutation test for comparing within Person 1 using the reference and

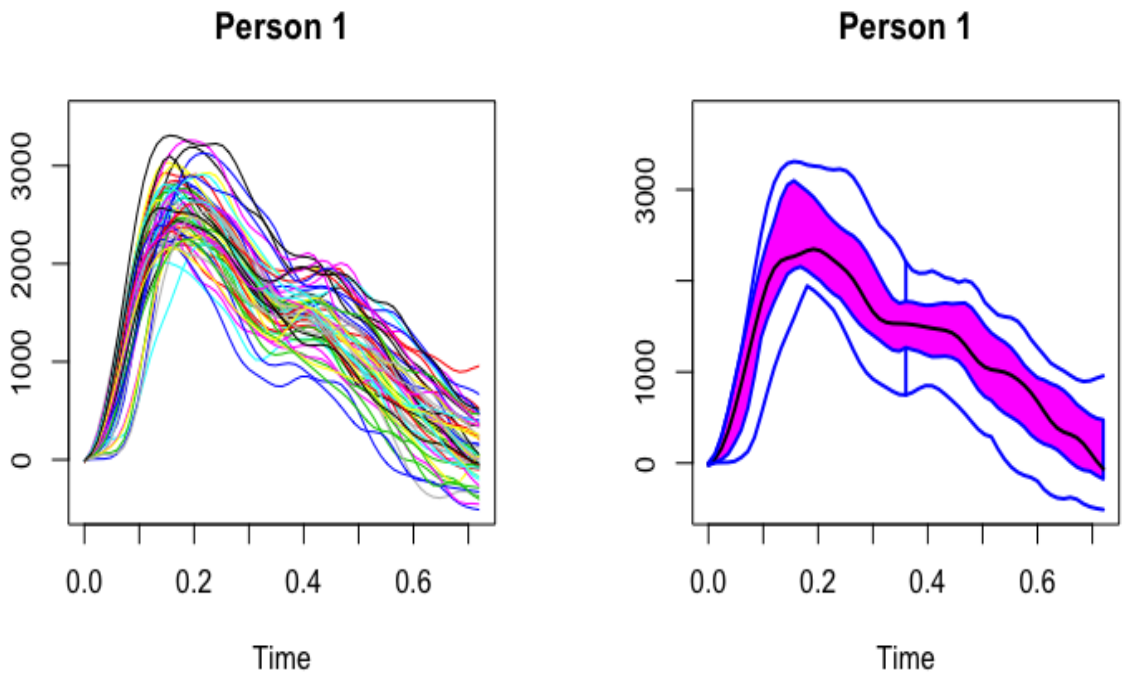

Figure 5.1: *The original curves and functional boxplot for Person 1.*

the rest group is 0.53. Thus, we do not have enough evidence to conclude that these two sets of curves come from two different people. To assure the identity, we further apply the equivalence testing using the nonparametric bootstrap procedure discussed in Section 3.3.2 to see if we can confirm that these two sets of pulse waves generated from the same person. According to the method for determining the evaluation range discussed in Section 4.3.2, the range of equivalence evaluation for Person 1 is $(0.06, 0.45)$. Therefore, we compare the confidence interval computed from the nonparametric bootstrap procedure with the equivalence bands in the range of $(0.06, 0.45)$. If the 90% confidence interval falls entirely within the equivalence bands between the time of 0.06 and 0.45, statistically significant equivalence can be concluded for the two groups of curves from Person 1.

Figure 5.2 shows the 90% confidence interval for the ratio of the population mean curves from the reference and the test group of Person 1. The equivalence bands are shown in red. We see that the confidence interval falls completely with the equivalence bands between the

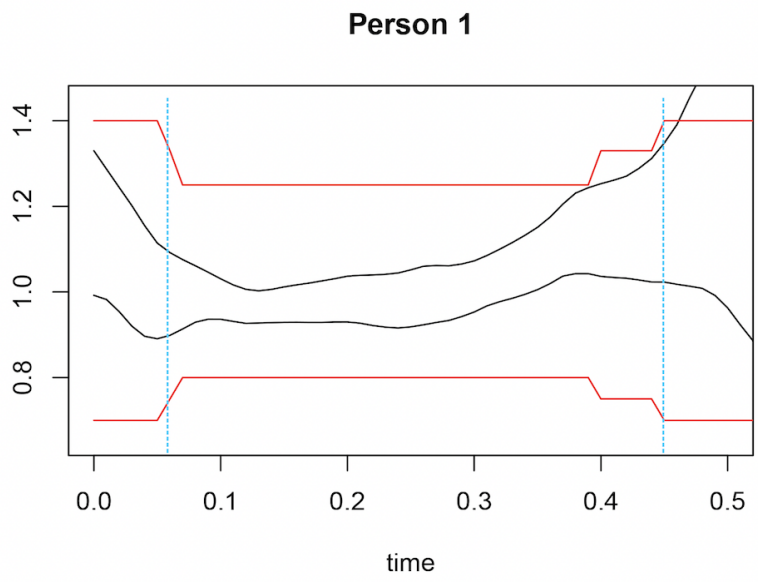

Figure 5.2: *90% confidence interval for the ratio of population mean curves from Person 1, with the equivalence bands shown in red and the evaluation range in blue.*

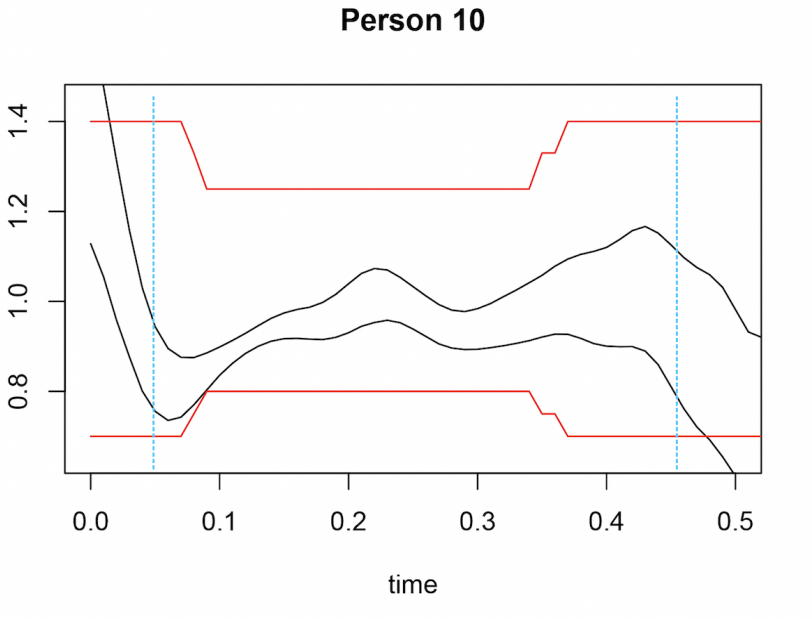

Figure 5.3: *90% confidence interval for the ratio of population mean curves from Person 10, with the equivalence bands shown in red and the evaluation range in blue.*

equivalence range of (0.06, 0.45). Therefore, we conclude that the reference and the test sets are generated from the same person. Namely, the identity is assured for Person 1. Similarly, we can also authenticate the identity of Person 10 from Figure 5.3. Figure 5.3 indicates the

90% confidence intervals for the ratio of the population mean curves from the reference and the rest group of Person 10, with the equivalence bands shown in red. The evaluation range computed according to Section 4.3.2 is (0.05, 0.46). Thus, we conclude that the reference and the test groups are from the same person since the confidence interval falls entirely within the equivalence limits in the evaluation range. However, from Figure 5.4, equivalence cannot be concluded for Person 2 and Person 7. Table 2.2 shows that the permutation p-value for comparing within Person 2 is 0.001. Therefore the first half and the second half curves are generated from different population means. It is not surprising that the 90% confidence interval falls outside the equivalence bands and we fail to confirm the identity of Person 10 from equivalence test. The permutation p-value for comparing within Person 7 is 0.32 as shown in Table 2.2. However, there is a lot of variability in curves from Person 7, and some go below zero at around $t = 0.4$. Moreover, several pulse waves from Person 7 do not seem to have the most appropriate starting points. Thus, the 90% confidence interval for the ratio of the population means dramatically widens at the beginning and around time 0.4.

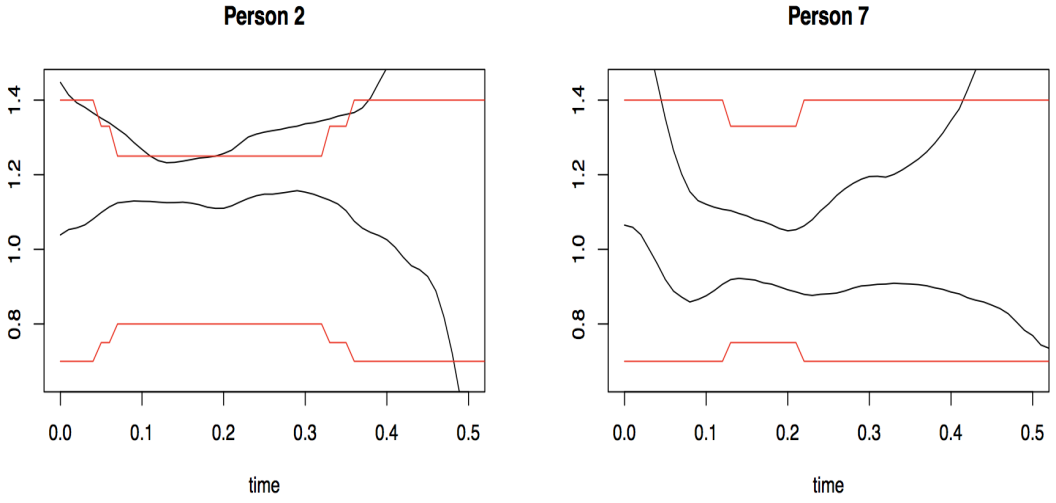

Figure 5.4: *90% confidence interval for the ratio of population mean curves from Person 2 and Person 7, with the equivalence bands shown in red.*

In the present work, we explored the use of pulse waves collected using PPG for identity

authentication. First, we visualized the pulse wave data using functional boxplots, which gave an assessment of the shape of the data and identified unusual observations. Functional boxplots also helped to adjust data by shifting pulse waves to a more appropriate starting point. Functional ANOVA and permutation tests were then employed to demonstrate that the identities in a small group of test subjects could be differentiated and compared by their pulse waveforms. We successfully differentiated the four persons in the sample through Functional ANOVA and permutation tests. The primary task of the project was to confirm the identity of a person, i.e., decide whether a given person, the test subject, is whom they claim to be, namely the reference subject. We developed a new equivalence testing procedure using nonparametric bootstrap, including the process of building the equivalence hypothesis and test statistics, determination of evaluation range and equivalence bands to determine whether the pulse waves of the test subject and the pulse waves of the reference subject were close enough to be considered equivalent. The identities of Person 1 and Person 10 were successfully confirmed through the nonparametric bootstrap procedure. However, we could not assure the identity of Person 2 and Person 7 due to issues such as large variabilities and negative values at the tail of pulse waves, improper starting points, unstable environment when the pulse waves were collected, poor settings of the pulse oximeters.

5.2 Summary of Contributions

The equivalence bands proposed in Section 4.4 are one of the main contributions in this work. Currently, equivalence limits of $(0.8, 1.25)$ are frequently used in the bioequivalence literature and also recommended by FDA. Instead of analyzing the individual profile as functional data, the profile is usually reduced to some scalar values by computing the pharmacokinetic characteristics, such as the area under the curve and the maximum concentration. For highly variable drugs, the limits could be widened to $(0.75, 1.33)$ or even to $(0.7, 1.43)$, depending on the within-subject CV. However, there is no literature or guidance in determining the equivalence limits for a functional measurement of similarity. We employed a simulation study to propose the proper EBs which are based on the point-wise coefficient of variances.

Another contribution we made is to propose the NBFET procedure in Section 3.3.2. Fogarty and Small (2014) considered the difference in mean curves and the ratio of variances as metrics for equivalence. They performed separate equivalence tests for these two measurements of equivalence. In our work, we used the ratio of mean curves as the test statistic and incorporate the variances into the determination of the equivalence bands. Therefore, both the mean curve and the variance are included in the equivalence test, and we do not lose information from either of them.

Moreover, we generated pulse wave-like curves from Brownian motion and demonstrated that Brownian motion can be used to simulate other functional data with ———particular patterns. To the best of our knowledge, there are no studies on the use of functional data generated from Brownian motion to perform identity authentication using an equivalence test or to determine the equivalence limits. The procedure to determine the evaluation range we used in Section 4.3.2 may be only applicable to pulse waveforms. However, similar ideas or steps could be employed to establish the range in an equivalence test for other examples of functional data.

5.3 Future Work

In this paper, we first visualize the pulse waves by functional boxplots and get a sense of behavioral patterns of the underlying mean curve. Then functional ANOVA was performed via a permutation test to differentiate people as well as an equivalence test to confirm the identity of persons. Functional ANOVA is described in Section 2.2 and permutation test in Section 2.3. Section 3.3.2 introduced the nonparametric bootstrap procedure for assessing equivalence, with the corresponding simulation results shown in Chapter 4. These are all frequentist methodologies, and similar approaches also exist under the Bayesian framework. For example, we could employ Bayesian functional ANOVA for differentiating the wave pat-

terns of subjects. Markov chain Monte Carlo methods could be implemented, but there are problems with both regarding convergence and computational time since the dataset in our project is large. One possible solution is to use the integrated nested Laplace approximation (INLA) so that very accurate approximations to the posterior marginals can be directly computed. The main benefit of INLA is computational; where MCMC methods run for hours and days, INLA provides precise estimates in seconds and minutes.

Fogarty and Small (2014) proposed to conduct the following three steps when using the Bayesian framework for equivalence testing:

1. Define an equivalence region through expert consultation.

2. Define a probability value, called it γ, such that if $P\{H_a|Data\} \geq \gamma$, equivalence may be suggested. Using the suggestions of Jeffries (1961), a value of $\gamma = 0.75$ or $\gamma = 0.95$ may be appropriate.

3. Specify prior distributions for the metrics of equivalence that are commensurate with the researchers' prior belief of the alternative being true relative to the null.

We suggested using functional measures of location to evaluate practical equivalence in the frequentist case. However, it is not required to carry out a TOST in the Bayesian paradigm. The Bayesian approach allows the computation of the posterior probabilities of our functional metrics of equivalence which can then be compared to specified range directly. In another word, we directly calculate $P\{H_a|Data\}$ for each set of the equivalence hypotheses. INLA could be used to get both the posterior and predictive distribution for the metrics of equivalence.

Bibliography

Carlos Alberola-López and Marcos Martín-Fernández. A simple test of equality of time series. *Signal processing*, 83(6):1343–1348, 2003.

Mohamad E Alnaeb, Nasser Alobaid, Alexander M Seifalian, Dimitri P Mikhailidis, and George Hamilton. Optical techniques in the assessment of peripheral arterial disease. *Current vascular pharmacology*, 5(1):53–59, 2007.

Alphan Altinok and Matthew Turk. Temporal integration for continuous multimodal biometrics. In *Proceedings of the Workshop on Multimodal User Authentication*, number 1. Citeseer, 2003.

Douglas G Altman and J Martin Bland. Statistics notes: Absence of evidence is not evidence of absence. *Bmj*, 311(7003):485, 1995.

Sharon Anderson and Walter W Hauck. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics-Theory and Methods*, 12(23):2663–2692, 1983.

H Harry Asada, Phillip Shaltis, Andrew Reisner, Sokwoo Rhee, and Reginald C Hutchinson. Mobile monitoring with wearable photoplethysmographic biosensors. *IEEE Engineering in Medicine and Biology Magazine*, 22(3):28–40, 2003.

Manuel Febrero Bande and Manuel Oviedo de la Fuente. Functional data analysis and utilities for statistical computing. http://www.jstatsoft.org/v51/i04/, 2016. R package version 1.3.0.

Shu-Di Bao, Yuan-Ting Zhang, and Lian-Feng Shen. Physiological signal based entity authentication for body area sensor networks and mobile healthcare systems. In *2005 IEEE*

Engineering in Medicine and Biology 27th Annual Conference, pages 2455–2458. IEEE, 2005.

Lawrence Barker, Henry Rolka, Deborah Rolka, and Cedric Brown. Equivalence testing for binomial random variables: which test to use? *The American Statistician*, 55(4):279–287, 2001.

Lawrence E Barker, Elizabeth T Luman, Mary M McCauley, and Susan Y Chu. Assessing equivalence: an alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology*, 156(11):1056–1061, 2002.

BP Barnett, DL Kraitchman, C Lauzon, CA Magee, P Walczak, WD Gilson, A Arepally, and JWM Bulte. Radiopaque alginate microcapsules for x-ray visualization and immunoprotection of cellular therapeutics. *Molecular pharmaceutics*, 3(5):531–538, 2006.

Brad P Barnett, Aravind Arepally, Parag V Karmarkar, Di Qian, Wesley D Gilson, Piotr Walczak, Valerie Howland, Leo Lawler, Cal Lauzon, Matthias Stuber, et al. Magnetic resonance–guided, real-time targeted delivery and imaging of magnetocapsules immunoprotecting pancreatic islet cells. *Nature medicine*, 13(8):986–991, 2007.

Sam Behseta and Robert E Kass. Testing equality of two functions using bars. *Statistics in medicine*, 24(22):3523–3534, 2005.

Francesco Bergadano, Daniele Gunetti, and Claudia Picardi. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security (TISSEC)*, 5(4):367–397, 2002.

Roger L Berger and Jason C Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

Francesco Beritelli and Salvatore Serrano. Biometric identification based on frequency analysis of cardiac sounds. *IEEE Transactions on Information Forensics and Security*, 2(3): 596–604, 2007.

Philippe C Besse, Hervé Cardot, and David B Stephenson. Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27(4):673–687, 2000.

Lena Biel, Ola Pettersson, Lennart Philipson, and Peter Wide. Ecg analysis: a new approach in human identification. *IEEE Transactions on Instrumentation and Measurement*, 50(3): 808–812, 2001.

Donald J Birkett. Generics-equal or not? *Australian Prescriber*, 26(4):85–7, 2003.

Graciela Boente and Ricardo Fraiman. Kernel-based functional principal components. *Statistics & probability letters*, 48(4):335–345, 2000.

Adrian Bowman and Adelchi Azzalini. Smoothing methods for nonparametric regression and density estimation. http://www.stats.gla.ac.uk/~adrian/sm, 2015. R package version 2.2-5.4.

Lawrence D Brown, JT Gene Hwang, and Axel Munk. An unbiased test for the bioequivalence problem. *The annals of Statistics*, pages 2345–2367, 1997.

Roberto Brunelli and Tomaso Poggio. Face recognition: Features versus templates. *IEEE transactions on pattern analysis and machine intelligence*, 15(10):1042–1052, 1993.

Attaullah Buriro, Bruno Crispo, Filippo Delfrari, and Konrad Wrona. Hold and sign: a novel behavioral biometrics for smartphone user authentication. In *Security and Privacy Workshops (SPW), 2016 IEEE*, pages 276–285. IEEE, 2016.

Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1):11–22, 1999.

CP Chua and C Heneghan. Continuous blood pressure monitoring using ecg and finger photoplethysmogram. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 5117–5120. IEEE, 2006.

Eric Chern-Pin Chua, Stephen J Redmond, Gary McDarby, and Conor Heneghan. Towards using photo-plethysmogram amplitude to measure blood pressure during sleep. *Annals of biomedical engineering*, 38(3):945–954, 2010.

John B Conway. *A course in functional analysis*, volume 96. Springer Science & Business Media, 2013.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics*, 30(2):285–300, 2002.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122, 2004.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. On the use of the bootstrap for estimating functions with functional data. *Computational statistics & data analysis*, 51 (2):1063–1074, 2006.

Julien Damon and Serge Guillas. The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics*, 13(7):759–774, 2002.

Jacques Dauxois, Alain Pousse, and Yves Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of multivariate analysis*, 12(1):136–154, 1982.

Barbara M Davit, Mei-Ling Chen, Dale P Conner, Sam H Haidar, Stephanie Kim, Christina H Lee, Robert A Lionberger, Fairouz T Makhlouf, Patrick E Nwakama, Devvrat T Patel, et al. Implementation of a reference-scaled average bioequivalence approach for highly variable generic drug products by the us food and drug administration. *The AAPS journal*, 14(4):915–924, 2012.

Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

Mohammad Omar Derawi, Claudia Nickel, Patrick Bours, and Christoph Busch. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on*, pages 306–311. IEEE, 2010.

Ilaria DiMatteo, Christopher R Genovese, and Robert E Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25, 2012.

Randall L Eubank. *Nonparametric regression and spline smoothing*. CRC press, 1999.

Jianqing Fan and Sheng-Kuei Lin. Test of significance when data are curves. *Journal of the American Statistical Association*, 93(443):1007–1021, 1998.

Conor P Farrington and Godfrey Manning. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in medicine*, 9(12):1447–1454, 1990.

Manuel Febrero, Pedro Galeano, and Wenceslao González-Manteiga. A functional analysis of nox levels: location and scale estimation and outlier detection. *Computational Statistics*, 22(3):411–427, 2007.

Frédéric Ferraty and Philippe Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17(4):545–564, 2002.

Frédéric Ferraty and Philippe Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1):161–173, 2003.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

Sue Finch, Geoff Cumming, and Neil Thomason. Reporting of statistical inference in the journal of applied psychology: Little evidence of reform. *Educational and Psychological Measurement*, 2001.

Colin B Fogarty and Dylan S Small. Equivalence testing for functional data with an application to comparing pulmonary function devices. *arXiv preprint arXiv:1407.5079*, 2014.

Ricardo Fraiman and Graciela Muniz. Trimmed means for functional data. *Test*, 10(2): 419–440, 2001.

Ricardo Fraiman, Jean Meloche, Luis A García-Escudero, Alfonso Gordaliza, Xuming He, Ricardo Maronna, Víctor J Yohai, Simon J Sheather, Joseph W McKean, Christopher G Small, et al. Multivariate l-estimation. *Test*, 8(2):255–317, 1999.

Jeff Goldsmith, Fabian Scheipl, and Lei Huang. Regression with functional data, 2018. R package version 2.14.0.

YY Gu and YT Zhang. Photoplethysmographic authentication through fuzzy logic. In *Biomedical Engineering, 2003. IEEE EMBS Asian-Pacific Conference on*, pages 136–137. IEEE, 2003.

YY Gu, Y Zhang, and YT Zhang. A novel biometric approach in human verification by photoplethysmographic signals. In *Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference on*, pages 13–14. IEEE, 2003.

Peter Hall and Ingrid Van Keilegom. Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, pages 1511–1531, 2007.

David J Hand. *Information generation: how data rules our world*. 2007.

Jeffrey D Hart and Thomas E Wehrly. Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, 81(396):1080–1088, 1986.

Walter W Hauck and Sharon Anderson. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, 12(1):83–91, 1984.

Takeyuki Hida. Brownian motion. In *Brownian Motion*, pages 44–113. Springer, 1980.

David B Hitchcock, James G Booth, and George Casella. The effect of pre-smoothing functional data on cluster analysis. *Journal of Statistical Computation and Simulation*, 77 (12):1043–1055, 2007.

Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

Myung-Hoe Huh. Equivalence testing as an alternative to significance testing. *J Korean Stat Soc*, 23:199–206, 1994.

Harold Jeffries. Theory of probability, 1961.

Harold E Jones and Nancy Bayley. The berkeley growth study. *Child development*, 12(2): 167–173, 1941.

Samuel Karlin. *A first course in stochastic processes*. Academic press, 2014.

Cari G Kaufman, Stephan R Sain, et al. Bayesian functional {ANOVA} modeling using gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149, 2010.

A Reşit Kavsaoğlu, Kemal Polat, and M Recep Bozkurt. A novel feature ranking algorithm for biometric recognition with ppg signals. *Computers in biology and medicine*, 49:1–14, 2014.

W Khalifa, A Salem, M Roushdy, and K Revett. A survey of eeg based user authentication schemes. In *Informatics and Systems (INFOS), 2012 8th International Conference on*, pages BIO–55. IEEE, 2012.

Regina Y Liu et al. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

N Locantore, JS Marron, DG Simpson, N Tripoli, JT Zhang, KL Cohen, Graciela Boente, Ricardo Fraiman, Babette Brumback, Christophe Croux, et al. Robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

Yolanda Muñoz Maldonado, Joan G Staniswalis, Louis N Irwin, and Donna Byers. A similarity analysis of curves. *Canadian Journal of Statistics*, 30(3):373–381, 2002.

Michael Meyners. Equivalence tests–a review. *Food quality and preference*, 26(2):231–245, 2012.

KK Midha, MJ Rawson, and JW Hubbard. The bioequivalence of highly variable drugs and drug products. *International Journal of Clinical Pharmacology & Therapeutics*, 43(10), 2005.

Fabian Monrose and Aviel D Rubin. Keystroke dynamics as a biometric for authentication. *Future Generation computer systems*, 16(4):351–359, 2000.

Willie Bosseau Murray and Patrick Anthony Foster. The peripheral pulse wave: information overlooked. *Journal of clinical monitoring*, 12(5):365–377, 1996.

Michael Negin, Thomas A Chmielewski, Marcos Salganicoff, UM von Seelen, PL Venetainer, and Guanghua G Zhang. An iris biometric system for public and personal use. *Computer*, 33(2):70–75, 2000.

Ikenna Odinaka, Po-Hsiang Lai, Alan D Kaplan, Joseph A O'Sullivan, Erik J Sirevaag, and John W Rohrbaugh. Ecg biometric recognition: A comparative analysis. *IEEE Transactions on Information Forensics and Security*, 7(6):1812–1824, 2012.

R Todd Ogden, Carl E Miller, Kunio Takezawa, and Seishi Ninomiya. Functional regression in crop lodging assessment with digital images. *Journal of Agricultural, Biological, and Environmental Statistics*, 7(3):389–402, 2002.

Hannu Oja. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327–332, 1983.

Victor M Panaretos, David Kraus, and John H Maddocks. Second-order comparison of gaussian random functions and the geometry of dna minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.

David F Parkhurst. Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation equivalence tests improve the logic of significance testing when demonstrating similarity is important, and reverse tests can help show that failure to reject a null hypothesis does not support that hypothesis. *Bioscience*, 51(12):1051–1057, 2001.

S Pezzulli and BW Silverman. Some properties of smoothed principal components analysis for functional data. *Computational Statistics*, 8:1–1, 1993.

CCY Poon, XF Teng, YM Wong, C Zhang, and YT Zhang. Changes in the photoplethysmogram waveform after exercise. In *Computer Architectures for Machine Perception, 2003 IEEE International Workshop on*, pages 115–118. IEEE, 2004.

R Core Team. R: A language and environment for statistical computing. http://www.R-project.org/, 2013.

J. O. Ramsay, Hadley Wickham, Spencer Graves, and Giles Hooker. Functional data analysis. http://www.functionaldata.org, 2015. R package version 2.4.4.

James O Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

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

Jim O Ramsay. Functional components of variation in handwriting. *Journal of the American Statistical Association*, 95(449):9–15, 2000.

JO Ramsay. When the data are functions. *Psychometrika*, 47(4):379–396, 1982.

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

Ashok Samal and Prasana A Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern recognition*, 25(1):65–77, 1992.

Donald J Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–680, 1987.

ERJ Seitsonen, IKJ Korhonen, MJ Van Gils, M Huiku, JMP Lötjönen, KT Korttila, and AM Yli-Hankala. Eeg spectral entropy, heart rate, photoplethysmography and motor responses to skin incision during sevoflurane anaesthesia. *Acta Anaesthesiologica Scandinavica*, 49(3):284–292, 2005.

Murray R Selwyn and Nancy R Hall. On bayesian methods for bioequivalence. *Biometrics*, pages 1103–1108, 1984.

Murray R Selwyn, Arthur P Dempster, and Nancy R Hall. A bayesian approach to bioequivalence for the 2 x 2 changeover design. *Biometrics*, pages 11–21, 1981.

Kirk H Shelley. Photoplethysmography: beyond the calculation of arterial oxygen saturation and heart rate. *Anesthesia & Analgesia*, 105(6):S31–S36, 2007.

Bernard W Silverman et al. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24, 1996.

BW Silverman, Luc Devroye, and Laszlo Gyorfi. Two books on density estimation, 1985.

Terence Sim, Sheng Zhang, Rajkumar Janakiraman, and Sandeep Kumar. Continuous verification using multimodal biometrics. *IEEE transactions on pattern analysis and machine intelligence*, 29(4):687–700, 2007.

K Singh. A notion of majority depth. *Unpublished document*, 1991.

Zdeňka Sitová, Jaroslav Šeděnka, Qing Yang, Ge Peng, Gang Zhou, Paolo Gasti, and Kiran S Balagani. Hmog: New behavioral biometric features for continuous authentication of smartphone users. *IEEE Transactions on Information Forensics and Security*, 11(5):877–892, 2016.

Petros Spachos, Jiexin Gao, and Dimitrios Hatzinakos. Feasibility study of photoplethysmographic signals for biometric identification. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2011.

Ying Sun and Marc G Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 2012.

RJ Tempelman. Experimental design and statistical methods for classical and bioequivalence hypothesis testing with an application to dairy nutrition studies. *Journal of Animal Science*, 82(13_suppl):E162–E172, 2004.

Heizo Tokutaka, Yoshio Maniwa, Eikou Gonda, Masashi Yamamoto, Toshiyuki Kakihara, Masahumi Kurata, Kikuo Fujimura, Li Shigang, and Masaaki Ohkita. Construction of a general physical condition judgment system using acceleration plethysmogram pulse-wave analysis. In *International Workshop on Self-Organizing Maps*, pages 307–315. Springer, 2009.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531, 1975.

Yehuda Vardi and Cun-Hui Zhang. The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Matt P Wand and M Chris Jones. *Kernel smoothing*. Chapman and Hall/CRC, 1994.

Shanshan Wang, Wolfgang Jank, Galit Shmueli, and Paul Smith. Modeling price dynamics in ebay auctions using differential equations. *Journal of the American Statistical Association*, 2012.

Stefan Wellek. *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press, 2010.

Wilfred J Westlake. Bioavailability and bioequivalence of pharmaceutical formulations. *Biopharmaceutical statistics for drug development*, pages 329–352, 1988.

WJ Westlake. Bioequivalence testing–a need to rethink. *Biometrics*, 37(3):589–594, 1981.

Hulin Wu and Jin-Ting Zhang. *Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches*, volume 515. John Wiley & Sons, 2006.

Jianchu Yao, Xiaodong Sun, and Yongbo Wan. A pilot study on using derivatives of photoplethysmographic signals as a biometric identifier. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4576–4579. IEEE, 2007.

Jin-Ting Zhang. *Analysis of variance for functional data*. Chapman and Hall/CRC, 2013.

Appendix A

Additional Simulation Results for Determining Equivalence Bands

This appendix shows the simulation results for determining the most appropriate EBs for different options of average CV.

	Ref Data	Test Data	mean of CV	EBs	# rejecting	Clopper-Pearson CI
H_a True	Mean E Var=25	Mean E Var=25	0.136	(0.8, 1.25)	1000/1000	(0.996, 1.000)
				(0.75, 1.33)	1000/1000	(0.996, 1.000)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean E Var=40	Mean E Var=40	0.215	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	922/1000	(0.904, 0.938)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean E Var=50	Mean E Var=50	0.337	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	636/1000	(0.605, 0.666)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean E Var=60	Mean E Var=60	0.411	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	68/1000	(0.0532, 0.0854)
				(0.7, 1.43)	984/1000	(0.974, 0.991)

Table A.1: *The null hypothesis of difference is false. The power for different choices of equivalence bands with different averages of CV.*

	Ref Data	Test Data	mean of CV	EBs	# rejecting	Clopper-Pearson CI
H_a True	Mean F Var=40	Mean F Var=40	0.141	(0.8, 1.25)	996/1000	(0.990, 0.999)
				(0.75, 1.33)	1000/1000	(0.996, 1.000)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean F Var=80	Mean F Var=80	0.239	(0.8, 1.25)	12/1000	(0.00621, 0.0209)
				(0.75, 1.33)	979/1000	(0.968, 0.987)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean F Var=100	Mean F Var=100	0.328	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	626/1000	(0.595, 0.656)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean F Var=120	Mean F Var=120	0.408	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	147/1000	(0.126, 0.170)
				(0.7, 1.43)	993/1000	(0.986, 0.997)
H_a True	Mean B Var=40	Mean C Var=50	0.125	(0.8, 1.25)	998/1000	(0.993, 0.999)
				(0.75, 1.33)	1000/1000	(0.996, 1.000)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean B Var=80	Mean C Var=80	0.218	(0.8, 1.25)	67/1000	(0.0523, 0.0843)
				(0.75, 1.33)	988/1000	(0.979, 0.994)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean B Var=100	Mean C Var=100	0.334	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	596/1000	(0.565, 0.627)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean B Var=120	Mean C Var=100	0.467	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	244/1000	(0.218, 0.272)
				(0.7, 1.43)	995/1000	(0.988, 0.998)
H_a True	Mean D Var=25	Mean E Var=25	0.162	(0.8, 1.25)	994/1000	(0.987, 0.999)
				(0.75, 1.33)	1000/1000	(0.996, 1.000)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean D Var=40	Mean E Var=40	0.264	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	826/1000	(0.801, 0.849)
				(0.7, 1.43)	1000/1000	(0.996, 1.000)
H_a True	Mean D Var=50	Mean E Var=50	0.352	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	320/1000	(0.291, 0.349)
				(0.7, 1.43)	925/1000	(0.906, 0.941)
H_a True	Mean D Var=60	Mean E Var=60	0.426	(0.8, 1.25)	0/1000	(0, 0.00368)
				(0.75, 1.33)	10/1000	(0.00481, 0.0183)
				(0.7, 1.43)	910/1000	(0.891, 0.927)

Table A.2: *Continued: the null hypothesis of difference is false. The power for different choices of equivalence bands with different averages of CV.*

Appendix B

Additional Simulation Results for Validation

	Ref Data	Test Data	average of CV	# rejecting	Clopper-Pearson CI
H_a True	Mean B Var=40	Mean B Var=40	0.125	889/1000	(0.868, 0.908)
H_a True	Mean B Var=80	Mean B Var=80	0.218	987/1000	(0.978, 0.993)
H_a True	Mean B Var=100	Mean B Var=100	0.334	989/1000	(0.980, 0.994)
H_a True	Mean B Var=120	Mean B Var=120	0.467	944/1000	(0.928, 0.957)
H_a True	Mean C Var=40	Mean C Var=40	0.172	937/1000	(0.920, 0.951)
H_a True	Mean C Var=80	Mean C Var=80	0.225	976/1000	(0.964, 0.985)
H_a True	Mean C Var=100	Mean C Var=100	0.313	986/1000	(0.977, 0.992)
H_a True	Mean C Var=120	Mean C Var=120	0.429	890/1000	(0.869, 0.909)

Table B.1: *The null hypothesis of difference is false. The power of the test for difference averages of CV based on the corrected equivalence bands.*

	Ref Data	Test Data	average of CV	# rejecting	Clopper-Pearson CI
H_a True	Mean D Var=25	Mean D Var=25	0.162	958/1000	(0.944, 0.970)
H_a True	Mean D Var=40	Mean D Var=40	0.264	976/1000	(0.964, 0.985)
H_a True	Mean D Var=50	Mean D Var=50	0.352	984/1000	(0.974, 0.991)
H_a True	Mean D Var=60	Mean D Var=60	0.426	922/1000	(0.904, 0.938)
H_a True	Mean E Var=25	Mean E Var=25	0.136	865/1000	(0.842, 0.886)
H_a True	Mean E Var=40	Mean E Var=40	0.245	914/1000	(0.895, 0.931)
H_a True	Mean E Var=50	Mean E Var=50	0.328	955/1000	(0.940, 0.967)
H_a True	Mean E Var=60	Mean E Var=60	0.416	886/1000	(0.865, 0.905)
H_a True	Mean F Var=40	F Var=40	0.141	828/1000	(0.803, 0.851)
H_a True	Mean F Var=80	Mean F Var=80	0.239	927/1000	(0.909, 0.942)
H_a True	Mean F Var=100	Mean F Var=100	0.328	951/1000	(0.936, 0.964)
H_a True	Mean F Var=120	Mean F Var=120	0.408	865/1000	(0.842, 0.886)

Table B.2: *Continued: the null hypothesis of difference is false. The power of the test for difference averages of CV based on the corrected equivalence bands.*

Appendix C

R Programs

```
########################################
## R version 3.5.0 (2018-04-23)
## Packages used:
## fda Version 2.4.8
## fda.usc Version 1.4.0
## sm Version 2.2-5.6
########################################
```

C.1 Functional Boxplots

```
## functional data analysis, such as functional boxplot
library(fda)
## Smoothing Methods for Nonparametric Regression and Density Estimation
library(sm)
## read the original curves for Person 1
person1<-read.csv("/Users/mengjiaowu/Documents/Research/ModifiedData
/1-cycles2.csv",header=FALSE)
## the number of cycles
ncurves<-0.5*dim(person1)[2]-1
```

```
## the matrix for time and PPG values
T1<-matrix(0,nrow=dim(person1)[1],ncurves)
W1<-matrix(0,nrow=dim(person1)[1],ncurves)
## make the start time for each curve be zero
for (i in 1:ncurves){
  T1[,i]<-person1[,2*i+1]-min(person1[,2*i+1],na.rm=TRUE)
  W1[,i]<-person1[,2*i+2]
}
## shift the curves so that they all start at zero
for (i in 1:ncurves){
  W1[,i]<-W1[,i]-W1[,i][1]
}
P1matrix<-matrix(0,nrow=ncurves,ncol=71)
## smooth the original curves
for (i in 1:ncurves){
  P1matrix[i,]<-sm.regression(na.omit(T1[,i]),na.omit(W1[,i]),h=0.01,
                              eval.points=seq(0,0.84,by=0.012))$estimate
}
par(mfrow=c(1,2))
## make functional boxplots for Person 1
fbplot(t(P1matrix[,1:61]),x=seq(0,0.72,by=0.012),method="MBD",
       ylim=c(-500,3800),xlim=c(0,0.72),
       fullout=TRUE,xlab="Time",ylab="")
## add the title for the functional boxplot
title(main="Person␣1")
## plot the original curves
plot(seq(0,0.72,by=0.012),P1matrix[1,][1:61],type="l",ylim=c(-500,3500),
       xlab="Time",ylab="",main="Person␣1")
for(i in 2:ncurves){
```

```
  lines(seq(0,0.72,by=0.012),P1matrix[i,][1:61],col=i)
}
```

C.2 Functional ANOVA

```
## comparing four persons using ANOVA
library(fda.usc) ## functional ANOVA
## P1matrix is smoothing curves after truncation for person 1
## and each column represent a single pulse wave
mydata<-rbind(P1matrix,P2matrix,P7matrix,P10matrix)
## compute the number of curves for each person
n1<-dim(P1matrix)[2]
n2<-dim(P2matrix)[2]
n7<-dim(P7matrix)[2]
n10<-dim(P10matrix)[2]
## assign the lables for each person
group<-rep(c(1,2,7,10),c(n1,n2,n7,n10))
## one-way functional ANOVA based on 1000 bootstrap samples
res<-anova.onefactor(fdata(mdata),group,nboot=1000,plot=TRUE)
## save the graph
quartz.save("fanova6person.pdf","pdf",width=9)
```

C.3 Permutation Test

```
## permutation test using smoothed and shifted data to compare Person 1
##and Person 7
##the number of permutations
N<-1000
## P1matrix and P7matrix contain the smoothed and shift data
```

```
## from Person 1 and Person 7
P1_7<-rbind(P1matrix[,1:61],P7matrix[,1:61])
## each permutation generated one value of difference in areas
Diff1_7<-rep(0,N)
## n1: the number of curves of Person 1
n1<-nrow(P1matrix)
for (s in 1:N){
  time<-seq(0,0.72,by=0.012)
  ## randomly select n1 curves without replacement
  sample<-sample(1:nrow(P1_7),n1)
  traintime<-seq(0,0.72,by=0.012)
  ## create the training/reference data
  traindata<-t(P1_7)[,sample]
  num1<-matrix(0,nrow=ncol(traindata),ncol=length(time))
  dem1<-matrix(0,nrow=ncol(traindata),ncol=length(time))
  ## compute the local linear regression estimate for the reference data
  for (j in 1:length(time)){
    for(i in 1:ncol(traindata)){
      R<-var(na.omit(traindata[,i]))*diag(length(na.omit(traintime)))
      weight<-chol2inv(chol(R))
      num1[i,j]<-t(dnorm(na.omit(traintime)-time[j],0,0.01))%*%(diag(weight)
                        *na.omit(traindata[,i]))
      dem1[i,j]<-t(dnorm(na.omit(traintime)-time[j],0,0.01))%*%diag(weight)
    }
  }
  Numtrain<-apply(num1,2,sum)
  Demtrain<-apply(dem1,2,sum)
  ## etatrain is the local linear regression estimate for the reference data
  etatrain<-Numtrain/Demtrain
```

```
  ######################################################################
  testtime<-seq(0,0.72,by=0.012)
  testdata<-t(P1_7)[,-sample]
  num1<-matrix(0,nrow=ncol(testdata),ncol=length(time))
  dem1<-matrix(0,nrow=ncol(testdata),ncol=length(time))
  ## compute the local linear regression estimate for the test data
  for (j in 1:length(time)){
    for(i in 1:ncol(testdata)){
      R<-var(na.omit(testdata[,i]))*diag(length(na.omit(testtime)))
      weight<-chol2inv(chol(R))
      num1[i,j]<-t(dnorm(na.omit(testtime)-time[j],0,0.01))%*%(diag(weight)
                      *na.omit(testdata[,i]))
      dem1[i,j]<-t(dnorm(na.omit(testtime)-time[j],0,0.01))%*%diag(weight)
    }
  }
  Numtest<-apply(num1,2,sum)
  Demtest<-apply(dem1,2,sum)
  ## etatest is the local linear regression estimate for the test data
  etatest<-Numtest/Demtest

  ## compute the difference between two estimated curves
  h<-etatrain-etatest
  tdiff<-0.012
  ## compute the area between two curves using trapzoid rule
  Diff1_7[s]<-sum(abs(h[-1]+h[-length(h)])*tdiff/2)
}
## We computed the difference between the area under the estimated curve
## from Person 1 and from Person 7 is 121.595
p_value=mean(Diff1_7>121.595)
```

C.4 Brownian motion

```
## generate the curves from Brownian motion
xx<-seq(0,1,by=0.01)
## the mean is a mixture of Beta densities
mean.vector<-500*dbeta(xx,2.1,12)+550*dbeta(xx,4,5)
## the variance for the Brownian motion
vari<-40
## number of realizations
nreal<-50
## number of grid points
np<-100
## generate a group of 50 curves from Brownian motion
## with the given mean and varaince
mydata<-matrix(0,nrow=nreal,ncol=length(mean.vector))
fit<-matrix(0,nrow=nreal,ncol=length(mean.vector)-1)
for (i in 1:nreal){
  ## generate values from normal distribution with mean 0 and variance=vari
  dis<-rnorm(np,0,vari)
  dis<-c(0,dis)
  ## compute the cumulatice sum
  dis<-cumsum(dis)
  ## curve generated from Brownian motion with a given mean and variance
  mydata[i,]<-mean.vector+dis
  ## apply the nonparametric regression smoothing
  fit[i,]<-sm.regression(xx,curve1[i,],h=0.025,ngrid=np)$estimate
}


par(mfrow=c(1,2))
```

```
## plot the simualted data
time<-seq(0,1,length=np)
plot(time,fit[1,],type="l",ylim=c(-800,2800),col=4,ylab="",
    main="Simulated curves after smoothing")
for (i in 2:nreal){
  lines(time,fit[i,],col=4)
}
## make a functional boxplot for the simulated data
fbplot(t(fit),x=seq(0,1,length=np),method="MBD",xlim=c(0,1))
title("Functional boxplot of the simulated curves")
```

C.5 Nonparametric Bootstrap

```
## Nonparametric bootstap for the ratio of mean curves
## First, we generate curves from Brownian motion with two given
## population means: mean.vector1 and mean.vector2
nreal<-50 ## 50 realizations for each mean
np<-100 ## the number of evaluation/grid points
xx<-seq(0,1,by=0.01)
mean.vector1<-500*dbeta(xx,2.1,12)+550*dbeta(xx,4,5)
mean.vector2<-500*dbeta(xx,2.6,12)+450*dbeta(xx,4,5)

## genetate curves from Brownian motion and store them in curve1 and curve 2
curve1<-matrix(0,nrow=nreal,ncol=length(mean.vector1))
curve2<-matrix(0,nrow=nread,ncol=length(mean.vector1))
## apply nonparametric regression (smoothing) to curve1 and curve2,
## and store the results in fit1 and fit2
fit1<-matrix(0,nrow=nreal,ncol=np)
fit2<-matrix(0,nrow=nreal,ncol=np)
```

```
for (i in 1:nreal){
  dis<-rnorm(np,0,vari1)
  dis<-c(0,dis)
  dis<-cumsum(dis)
  curve1[i,]<-mean.vector1+dis
  fit1[i,]<-sm.regression(xx,curve1[i,],h=0.025,ngrid=np)$estimate
}
for (i in 1:nreal){
  dis<-rnorm(np,0,vari2)
  dis<-c(0,dis)
  dis<-cumsum(dis)
  curve2[i,]<-mean.vector2+dis
  fit2[i,]<-sm.regression(xx,curve2[i,],h=0.025,ngrid=np)$estimate
}

## Then, we apply the nonparametric bootstrap to fit1 and fit2
## the number of bootstrap samples
bsample<-1000
## the number of iterations
niter<-500
## Ratio stores the ratios of the estimated curves of the
## reference and the test group from each iteration
Ratio<-matrix(0,nrow=bsample,ncol=np)
## equivalence indicates if we reject the null of difference
## in each iteration
equivalence<-rep(-1,niter)
for (i in 1:niter){
  smRef_1<-matrix(0,nrow=bsample,ncol=np)
  smTest_1<-matrix(0,nrow=bsample,ncol=np)
```

```
  for (k in 1:bsample){
    Time<-seq(0,1,length=np)
    ## draw bootstrap samples for the reference group
    Cref<-sample(1:nreal,nreal,replace=TRUE)
    dataRef<-fit1[Cref,]
    ## draw bootstrap samples for the test group
    Ctest<-sample(1:nreal,nreal,replace=TRUE)
    dataTest<-fit2[Ctest,]
    ## compute the nonparametric regression estimate for repeated measurement
    ## curves for both the reference and the test group
    smRef_1[k,]<-sm.rm(Time,dataRef,minh=0.01,maxh=0.01,display="none")
                          $aux$mean
    smTest_1[k,]<-sm.rm(Time,dataTest,minh=0.01,maxh=0.01,display="none")
                         $aux$mean
  }

  ## compute the ratio of the estimated curves
  for (k in 1:bsample){
    Ratio[k,]<-smTest_1[k,]/smRef_1[k,]
  }
  ## compute the 90% confidence interval from bootstrapping results
  ratio_statqt<-apply(thetaRatio,2,function(x)
                                {quantile(x,probs=c(0.05,0.95),type=7)})
  ## if the confidence interval falls entirely within the equivalence limits,
  ## we conclude the equivalence
  equivalence[i]<-ifelse(min(ratio_statqt[1,])> 0.8
                           & max(ratio_statqt[2,])<1.25,1,0)
}
```