5-2019

Sampling Studies for Longitudinal Functional Data

Toni Jassel
Montclair State University

ABSTRACT

We study the data setting consisting of functional data sets repeatedly observed over time. The focus is on the dynamic prediction of the future trajectory for a subject. Regression methods based on dynamic functional models are used for dynamic prediction of individual trajectories. We propose strategies for the selection of the study sampling design in the context of longitudinal functional data. An application to simulated child growth data is presented. The height-for-age z-score (HAZ) was the response variable in the functional dynamic models for prediction. The intent was to recommend four months for removal in our initial historic data set. We quantify the effect on dynamic prediction performance when several data missing scenarios and methods of data imputation were considered. The effectiveness of seven methods of data imputation in the setting of longitudinal functional data were examined.

MONTCLAIR STATE UNIVERSITY

Sampling Studies for Longitudinal Functional Data

by

Toni Jassel

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

May 2019

College/School Science and Mathematics

Department Mathematical Sciences

Thesis Committee:

[redacted]

Dr. A Vanlescu

Thesis Sponsor

[redacted]

Dr. Andrew McDougall

Committee Member

[redacted]

Dr. Haiyan Su

Committee Member

SAMPLING STUDIES FOR LONGITUDINAL FUNCTIONAL DATA

A Thesis

Submitted in partial fulfillment of the requirements

For the degree of Master of Science

By

TONI JASSEL

Montclair State University

Montclair, NJ

2019

Acknowledgement

I would first like to thank my thesis advisor Dr. Andrada E. Ivanescu for her unwavering patience, support, and guidance throughout this journey.

I would like to acknowledge the members of my thesis committee, Dr. Andrew McDougall and Dr. Haiyan Su. I am incredibly appreciative of their review, and I am gratefully indebted to them both for their valuable insight on this thesis.

I would also like to thank two additional faculty members at Montclair State University, Dr. Eric Forgoston and Ms. Chrystel Williams, without whom I would not have found the courage to pursue this opportunity.

Last but not least, I must express my sincerest gratitude to: my father for inspiring me to dream big, my mother for encouraging me to be the best version of myself, and my brother and sister for providing me with continuous love and support throughout my years of study. This accomplishment would not have been possible without them.

Thank you.

Table of Contents

Chapter 1: Introduction 1

Chapter 2: Literature Review 6
2.1 Longitudinal Data 6
2.2 Functional Data 7
2.3 Longitudinal Functional Data 9

Chapter 3: Analysis of functional data 12
3.1 Observed Functional Data 12
3.2 Sample Mean, Standard Deviation, and Covariance 13
3.3 Functional Principal Component Analysis 14
3.4 Fragment 1 HAZ Data Set 15

Chapter 4: Methods for Imputation 18
4.1 Subject Specific Regression 19
4.2 Linear Inter and Extrapolation 20
4.3 Last Observation Carried Forward (LOCF) 20
4.4 Linear Mixed Effects Model 20
4.5 Functional Principal Components Analysis 21
4.6 Penalized Smoothing Method 1 23
4.7 Penalized Smoothing Method 2 25

Chapter 5: Methods of Dynamic Prediction 27
5.1 Benchmark Dynamic Method (BENDY) 28
5.2 Dynamic Linear Model (DLM) 29
5.3 Dynamic Penalized Function Regression (DPFR) 30
5.4 Dynamic Penalized Function-on-function Regression (DPFFR) 30

Chapter 6: Numerical Study 32
6.1 Data Generation 33
6.2 Simulation of Missing Data 34
6.3 Metrics 35
6.4 Results for Dynamic Prediction 35

Chapter 7: Dynamic Prediction for Longitudinal Functional Data 43
7.1 HAZ Data 43
7.2 Methods of Dynamic Prediction 43

7.3 Results 44

Chapter 8: Conclusion 46

List of Figures

Figure 1: The graph shows fragment one HAZ data. This includes 16 data points for 197 subjects spanning their first fifteen months of life. Each line represents a subject, and each point represents the respective HAZ measurement taken at the indicated month.

Figure 2: The graphs depicts the fragment one HAZ data for patients 23, 56, and 112 as well as a graphical comparison of the three subjects.

Figure 3: This graph depicts the functional principal components fit for subjects 23, 56, and 112. The gray points show the observed values for the given subjects. The solid black lines display the fPCA fit. This is based upon the availability of the complete fragment 1 data set.

Figure 4: The above graphs are two examples of B- splines similar to the ones we use in our first method of penalized smoothing PLS1.

Figure 5: The above graphs are two examples of Fourier basis functions similar to the ones we use in our second method of penalized smoothing.

Figure 6: The above graph depicts the fragment 1 and fragment 2 data for all 197 subjects. The model used was the DPFFR model. Highlighted are subjects 23, 56, and 112.

Figure 7: Shown is the beta function employed in the DPFFR model for data generation.

Figure 8: Shown is the HAZ data for 3 subjects post removal of months 4, 11, 12, and 13. The gray points show the location of suggested HAZ data that are suggested for removal in the sampling study.

Chapter 1. Introduction

Functional data has been observed in many disciplines including medicine (Sorensen et al. 2013, Yao et al. 2005a, Yao et al. 2005b, Goldsmith et al. 2013, Xiao et al. 2016), environmental studies (Ramsay and Silverman 2005, Kokoszka and Reimherr 2017), biology and biomedical studies (Yao et al. 2005a, Ramsay and Silverman 2005, Leroux et al. 2018, Ieva and Paganoni 2016), and business (Goldberg et al. 2014, Shang 2017, Fan et al. 2014). The most common examples of functional data include weather and stock data, where data points are densely collected across the domain. That is, there exists a full set of functional observations for each subject in the study. Functional data is observed in the form of a sample of curves, where measurements for each curve are taken at discrete points on a given domain.

Longitudinal functional data arises when repeated functional data samples are observed. Longitudinal data tends to be more sparsely or irregularly spaced than functional data. In this work we study the data setting consisting of functional data samples repeatedly observed over time. This type of data is called longitudinal functional data (Park and Staicu 2015, Islam et al. 2016, Goldsmith et al. 2012, Chen et al. 2016). The focus is on the dynamic prediction of the future trajectory of a subject and we use dynamic functional models (Ivanescu et al. 2017) for dynamic prediction. Dynamic prediction (Goldberg et al. 2014, Chiou 2012, Ivanescu et al. 2017, Leroux et al. 2018) for functional data analysis is an active research area. We propose strategies for the selection of the sampling design for longitudinal functional data analysis from the point of view of dynamic prediction performance.

There are specific data sampling problems due to missing data that can be exhibited in a variety of functional data sets. Such sampling problems arise due to the time intervals at which the data was recorded, either regular (or dense) or irregular appointments when recording data. It is costly to record and obtain biomedical data values monthly for large populations. We propose to discuss strategies for selecting a sample of observation points that can be considered as candidates for removal when researchers need to gain some financial relief associated with sampling design for studies involving functional data.

First, consider all examples of dense functional data, that is, data that is sampled for many points throughout each day for each subject (such as Microsoft's stock prices that are recorded on a minute to minute basis, see, for example, Kokoszka and Reimherr 2017). Now, consider the case of sparsely sampled functional data, where, for example, growth data are sparsely sampled. That is, there are few sampling points that are spread differently across children (such as, one child's height-for-age z-score HAZ is measured at month 3 and 6 and another child's HAZ is measured at birth and month 2). A standard approach is to estimate the trajectory of every subject i at all data points, such as $\{Y_{i,l}; i = 1,2,\dots,n;\ \ l = 0,1,\dots,15\}$ based on the data available (Goldsmith et al. 2013). This approach considers functional principal components analysis (fPCA) for functional data. In addition to fPCA, we investigate several other methods for data imputation, including last observation carried forward, linear interpolation, and other methods discussed in Chapter 4, and apply these in several data scenarios and compare the different approaches with respect to prediction accuracy.

The research we conduct involves the study of longitudinal functional data (Goldsmith et al. 2012, Park and Staicu 2015). We use a data set for Height-for-Age Z-

scores (HAZ) (Ivanescu et al. 2017) that contains some simulated data for 197 patients for months 0-15, where month 0 corresponds to the month of birth. We employ methods of dynamic prediction to estimate the trajectory of the subject-specific HAZ curve for the future months 30-45. Our goal was to predict subject data for the second fragment of data (months 30-45) when using the first fragment of data (months 0-15) as predictive information. The objective is under the scope of dynamic prediction (Ivanescu et al. 2017) for functional data analysis. We also studied the effect of removal and imputing of the data of a specific month in the past. We study the effect of removing each of the inner data points (1-14) within our historic data set (0-15). The strategy we propose involves missing data imputation, followed by studying the effects of data imputation on dynamic prediction. Missing data was simulated at given months for historic data collected in fragment 1 (months 0-15). Missing historic data was dealt with by imputing a HAZ value where missing values occurred. Several methods were used for data imputation. Imputation methods carried out included linear interpolation and extrapolation (LIE), last observation carried forward (LOCF), functional principal components analysis (fPCA), linear mixed effects (LME), and penalized least squares (PLS). Imputation methods are presented in Chapter 4. After data was imputed for a specific month where data was simulated as missing, the future HAZ trajectory for months 30-45 was predicted using several dynamic prediction methods (Ivanescu et al. 2017). This process of studying data imputation for each month in the history of the HAZ process was undertaken at each specific inner month in the history of HAZ growth for months 1-14. We attempted to determine which months were the least problematic to remove when using the mean squared error for dynamic prediction as the metric. The month where data imputation yielded the smallest prediction

error was considered a signal that the month's data may be considered for removal, compared to other months where prediction error was larger. We used methods of dynamic prediction for functional data, such as dynamic function-on-function regression (DPFFR) (Ivanescu et al. 2017). We intend to find several months (4 months) in the history of the HAZ process for months 0-15 that can be considered for removal in a strategy to develop a sampling schedule with a smaller number of patient visits or appointments based on performance of dynamic prediction for the future trajectory.

Optimal sampling schedules is a recent area of research in functional data analysis. There are a number of different approaches that researchers have taken in developing these. Some propose a method of prediction to recover individual functions (Wu et al. 2018). Others focus on both predicting scalar outcomes and recovering individual functions (Ji and Muller 2017). Recently, Park et al. (2018) presented the concept of developing an optimal design strategy for both scalar outcomes and recovering individual functions simultaneously. While the above papers deal with optimizing sampling schedules for different functional data sets, they do not study the effects of differing methods of imputation. We propose an option for sampling a schedule design that incorporates the study of imputation methods for any missing data and the impact on dynamic prediction.

We study the use of dynamic prediction in the context of longitudinal functional data. This research we conduct will be useful in the medical field, highlighting the most and least critical times for sampling, and their effects on the overall dynamic prediction for the future of the trajectories in the data studied. This research has the potential to conserve fiscal resources in terms of scheduling the data acquisition framework. This research will focus on implementations that use R Statistical Software (R Core Team 2019).

In summary, our strategy includes simulating missing data, imputing values for the missing data, and employing a method of dynamic prediction. For our first step, we will simulate missingness in our data set one month at a time, starting with month 1 and ending with month 14. Secondly, we will impute a value for each missing data point using each of our imputation methods. Lastly, we will employ our methods of dynamic prediction to predict the future curve and compare the prediction performance across methods and data sampling designs.

Chapter 2. Literature Review

In this chapter, we provide an introductory presentation on several types of data sets discussed in the literature and related to longitudinal functional data.

2.1 Longitudinal Data

Data is longitudinal if it tracks the same information on a set of subjects over a period of time. The analysis of longitudinal data studies associations that change dynamically (Hedeker et al. 2006).

There are several types of longitudinal data observed in biology and biomedical studies. For example, in Yao et al. (2005a), a dataset from the Multicenter AIDS Cohort Study was examined. This dataset recorded the repeated measurements of physical exams, laboratory results, and CD4+ percentages for 283 subjects. CD4+ is a type of white blood cell that fights infection. Each subject had on average 8 observations between years 1984 and 1991.

Methods of working efficiently with longitudinal data are discussed in Wu et al. (2018) as they study two different data sets in order to derive an ideal sampling schedule for each dataset. They examine data for salivary cortisol, a stress biomarker that follows a non-linear profile. They also study urinary progesterone. The goal was to identify times during the day to collect salivary cortisol, and identify which days during the menstrual cycle to measure the urinary progesterone. They selected several sampling schedules.

In Diggle et al. (2002), a number of additional applications of longitudinal data analysis in the field of health sciences are discussed. A study examines the CD4+ cell count in a group of 369 men who tested HIV+. This study included a total of 2,376 measured against time since seroconversion (the time at which HIV was detected in the patient).

There was an average of 6 observations per patient over the course of 10 years. Another application shows how Sitka spruce trees grow against time. This study includes 27 subjects (trees), and the response variable is log (tree height * tree diameter squared). In this case, the study is done over 522 days, and each tree has exactly 13 observations. Two additional data sets referenced display seizure counts for epileptic patients and protein content in a cow's milk. Data is collected on a similar weekly schedule for both of these studies.

In Hedeker et al. (2006), a longitudinal study that exhibits incomplete data is discussed. The study includes 66 depressed inpatients. The response variable depression severity is measured using the Hamilton Depression Rating Scale. These measurements are taken weekly for six weeks. Only 46 patients had complete data for all time points. They had a total of 375 data points. Hedeker discusses mixed-effects regression models which make use of all available data points by allowing the intercept and time trends to vary for each subject.

All of the above examples reflect real world instances of longitudinal data analysis. Methods to address longitudinal data include mixed effects models. The methods of analysis referenced in these studies are similar to those we will examine later in the study.

Methods used for longitudinal data analysis include mixed effects models (Ruppert et al. 2003) and principal components analysis (Yao et al. 2005a). Some functional data methods can also be applied to longitudinal data (Ramsay et al. 2005).

2.2 Functional Data

Functional data is any data that is seen to vary over a continuum. Typically, this data is densely collected over a condensed time period, but the domain is not restricted to

time alone. These measurements typically follow highly volatile variables. That is, variables that can exhibit extreme changes over short periods of time. This is why we often see densely collected measurements over condensed time domains. A common example of this includes stock prices, and weather data, such as in Ramsay and Silverman (2005). In Sorensen et al. (2013), the similarities and differences between functional and longitudinal data are discussed. They are alike in that the data consist of repeated measurements for each subject. A difference is that longitudinal data often models expected value as polynomials or simple non-linear functions, whereas there is heightened flexibility in functional approaches. Moreover, functional data is considered to have a larger number of observed values for each subject.

Ramsay et al. (2005) presents a number of functional data examples. In one biomechanical example, force exerted by the thumb and forefinger is studied. Force was sampled at a rate of 500 times per second, but limited to a time interval that ranges from 0.0 to 0.30 seconds. The data set included 20 different recordings or curves.

Leroux et al. (2018) depicts a functional child growth data set. The data consists of 215 children, and 547 unique observations occurring between months 0 and 24 where month 0 represents birth. Each child had on average 34 measurements, where a majority of measurements took place during the first few months of the study. The study used length for age z-scores, weight for length z-scores, and weight for age z-scores to study the impact of Helicobacter pylori, bacterial infection, on child growth. These z-scores are metrics commonly used to model child growth on a scale relative to the World Health Organization's defined standard. Ieva and Paganoni (2016) present another example of a

functional data set that consists of ECG signals (noisy signals that describe the heart dynamics of each patient) for 149 subjects.

An application of functional data analysis is presented (Shang 2017) where the study is centered around the expected value of stock return. This is a particularly interesting problem as predictions were to be made based on dense data that spans very short time periods, for example, intraday returns. Using functional principal components analysis, the work in Shang (2017) attempts to forecast this functional time series.

Some popular methods for analyzing functional data include: functional principal components analysis, fPCA (Yao et al. 2005a, Goldsmith et al. 2013, Wrobel et al. 2016) and functional regression (Ramsay and Silverman 2005, Kokoszka and Reimherr 2017).

2.3 Longitudinal Functional Data

Data sets which consist of functional data that follow a longitudinal design are considered longitudinal functional data. While longitudinal responses used to solely follow scalar observations, new technologies allow the collection of functional observations (Goldsmith et al. 2012). Moving away from scalar responses, longitudinal functional data is of the form where functional data sets are now collected at multiple times over a longitudinal continuum. In recent studies (Park and Staicu 2015, Islam et al. 2016, Goldsmith et al. 2012), there has been an increased focus on prediction methods within longitudinal functional data studies.

In one paper, CCA-FA profiles, which reflect the fractional anisotropy (FA) along the corpous collosum (CCA), are collected using diffusion tensor imagining (DTI) from multiple sclerosis (MS) patients amongst multiple hospital visits giving rise to longitudinal functional data (Park and Staicu 2015). CCA relates to corpus callosum, a band of nerve

fibers that enables communication between the right and left hemispheres of the brain. They monitored the change in the CCA-FA profile to track the progression of the MS in the respective patients. The purpose of their work was to develop a model for predicting the full CCA-FA trajectory for any future visit that accounts for all of the dependence sources in the data. This paper also presents an example where physical activity count profiles are observed for a number of subjects over several consecutive days, and a secondary example where modality profiles are observed for MS patients across several hospital visits. The data examined is considered longitudinal functional data because it consists of functional observations of many subjects observed across multiple hospital visits over a period of time.

Islam et al. (2016) tracks the association between feed intake of lactating sows, and minute-by-minute relative humidity throughout the first 21 days of their lactating period. They use the data collected to study the efficiency of their longitudinal dynamic functional regression (LDFR) prediction model. This data is considered longitudinal functional data because the feed intake changes with respect to relative humidity, and this functional data is tracked for many lactating sows over a 21 day longitudinal period of time.

Chen et. al (2016) presents a longitudinal functional data study that makes use of imputation methods. The study includes subjects between the ages of 60 and 90 and aims to reveal daily activity patterns associated with human aging. Each subject was fitted with an Actiheart activity monitor that measured accelerometry (a measure of acceleration) counts every minute for a period of 7 days following a clinical visit. This data was collected over a number of clinical visits. For cases of missing values, data was imputed with the average measurement across all available days. Missingness arose during instances where

the subject was unable to wear the activity monitor. The study does not explore the effects of differing imputation methods on prediction.

Another study explores the relationship between cerebral white matter tracks in MS patients and their cognitive impairment over time (Goldsmith et al. 2012). MS results in lesions in white matter tracts, and thus leads to severe disabilities in patients. This study aims to develop a greater understanding of the relationship between MS and the resulting disabilities. The data set studies approximately 100 patients across a number of visits. Each patient has between 3 and 8 visits recorded. The density of the data in conjunction with the longitudinal observations result in its classification as longitudinal functional data.

Chapter 3 Analysis of Functional Data

3.1 Observed functional data

The most often used notation to depict a functional data sample is of the form Y_i , where i denotes the sample curve index, where i = 1, 2, …, n and n is the total number of curves. For functional data observed at discrete time points we use the notation $Y_{i,l}$ to represent the value of the HAZ curve for child i at month l. In this chapter, we will represent our functional data observed for fragment one data at the time interval from month 0 to month 15 with l={0, 1, …, 15}. Thus, the HAZ data for fragment 1 for n=197 subjects is denoted by $Y_{i,l}$ where l= {0, 1,…, 15}, i = 1,2,…,197.

The functional dataset may consist of n sampled curves observed at equally spaced time points over the interval $[0,15]$. It is typically not feasible that the values of Y will be known at all points in the continuous domain [0,15]. In a typical experiment or study, they will only be available at some selected points, and the points can vary for each curve Y_i (Kokoszka and Reimherr 2017). Functional data examples exist where the number of points observed are seen to range from small to large, and even instances where the number of points observed differs for each subject.

In terms of the HAZ data set, we use notation $Y_{i,l}$ to denote data for child i at month l. We can refer to our HAZ data set as densely sampled at each month. For each child the HAZ score exists for all points, however, for the data collection study the HAZ score was observed at selected times. In this particular study, HAZ scores are measured monthly for each subject.

3.2 Sample Mean, Standard Deviation, and Covariance

Once the data has been collected and imported into software, we can apply simple summary statistics using the raw functional data. Some metrics used to summarize functional data include the pointwise mean, pointwise standard deviation, and sample covariance function. The pointwise mean provides an initial estimate for the true functional mean and is computed as the average across the observed Y values (HAZ score values) over all subjects *n*. The computation for the sample mean is executed pointwise (Ramsay and Silverman 2005, Chapter 2, page 22) using the formula

$$\overline{Y}_l = \frac{1}{n}\sum_{i=1}^{n} Y_{i,l} \quad (1)$$

In addition to a measure of center, a measure of variability can also be computed. The pointwise standard deviation provides us with an estimate for the level of variability between the curves at any point l, such as using the formula below.

$$SD_l = \left\{\frac{1}{n-1}\sum_{i=1}^{n}(Y_{i,l} - \overline{Y}_l)^2\right\}^{1/2} \quad (2)$$

In relation to our data setting, the pointwise standard deviation quantifies variability of HAZ growth curves in relation to the pointwise mean calculated for all subjects.

The sample covariance function shows us the variability between all curves at two different time points l and s (Ramsay and Silverman 2005, Kokoszka and Reimherr 2017). The calculation (Ramsay and Silverman 2005, Chapter 2, page 22) is provided as follows.

$$\hat{c}_{l,s} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{i,l} - \overline{Y}_l)(Y_{i,s} - \overline{Y}_s) \quad (3)$$

There are some functional versions available for estimation of the true population parameters. For example, Bunea et al. (2011) presents a functional mean estimation technique for functional data. For the covariance estimation, several additional algorithms (Yao et al. 2005a, Xiao et al. 2016) are available.

3.3 Functional Principal Component Analysis (fPCA)

Estimated Functional Principal Components (fPC) are used to estimate deviations from the mean in a functional dataset. The goal is to find eigenfunctions $\hat{v}_j$ that reflect the most pertinent patterns of deviation from the mean function of the sample of curves. The eigenfunctions are computed from the estimated covariance matrix $\hat{c}$. We use existing implementations for computing the fPCA eigenfunctions (Goldsmith et al. 2013, Goldsmith et al. 2019). The term $\hat{\xi}_{i,j}$ is known as the score of sampled curve *i*, and represents how much of its shape can be attributed to the function $\hat{v}_j$ using the expression

$$Y_{i,l} - \overline{Y}_l \approx \sum_{j=1}^{P} \hat{\xi}_{i,j} \hat{v}_j(l). \tag{4}$$

The functional principal components have the property that $\hat{v}_1$ represents the most important deviation from the mean, $\hat{v}_2$ represents the second most important deviation. The total variability of a number of curves can be explained by the sum of explained variance attributed to the fPC's, $\hat{v}_j$. The percentage of variability that each function explains is related to the scores $\hat{\xi}_{i,j}$ (Kokoszka and Reimherr 2017, pg. 41). The larger the variance of the score, the larger the percentage, and thus the greater importance. In many cases, the first few functions account for the vast majority of the data variability. The computing of the functional principal components is derived from the covariance matrix, such as given

by equation (3), or other covariance surface estimation algorithms (Goldsmith et al. 2013, Xiao et al. 2016). We apply fPCA using the `fpca.sc` function within the refund (Goldsmith et al. 2019) R package.

3.4 Fragment 1 HAZ Data Set

We discuss in this section details about the functional dataset we work with in our analyses. Figure 1 is an illustration for HAZ data for fragment 1, ranging from month 0 to month 15. In the numerical study of Chapter 6, we discuss the HAZ data for fragment 2 which we simulate from fragment 1 data.

From this depiction, it is clear that we are dealing with a noisy data set. We are able to observe from Figure 1 that all of the children have observed HAZ scores between -4 and 3. The scores are seen to fluctuate between visits.

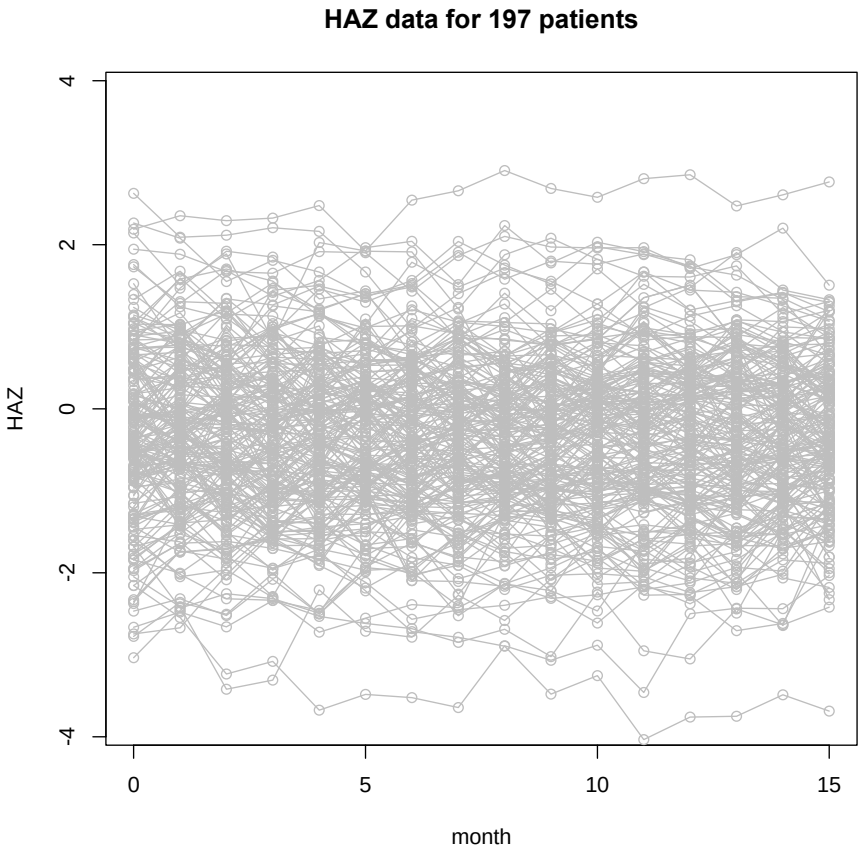


Figure 1: The graph shows fragment one HAZ data. This includes 16 data points for 197 subjects spanning their first fifteen months of life. Each line represents a subject, and each point represents the respective HAZ measurement taken at the indicated month.

Figure 2 shows HAZ data trajectories for three children. When visualizing the data based on individual subjects, we can see a more clear image of how the HAZ scores vary across time, in this case, months (0-15). Two of these subjects, (patient 23 and patient 112) are shown to follow similar trends in variation, with a steep increase, and then decline between months 13-15.

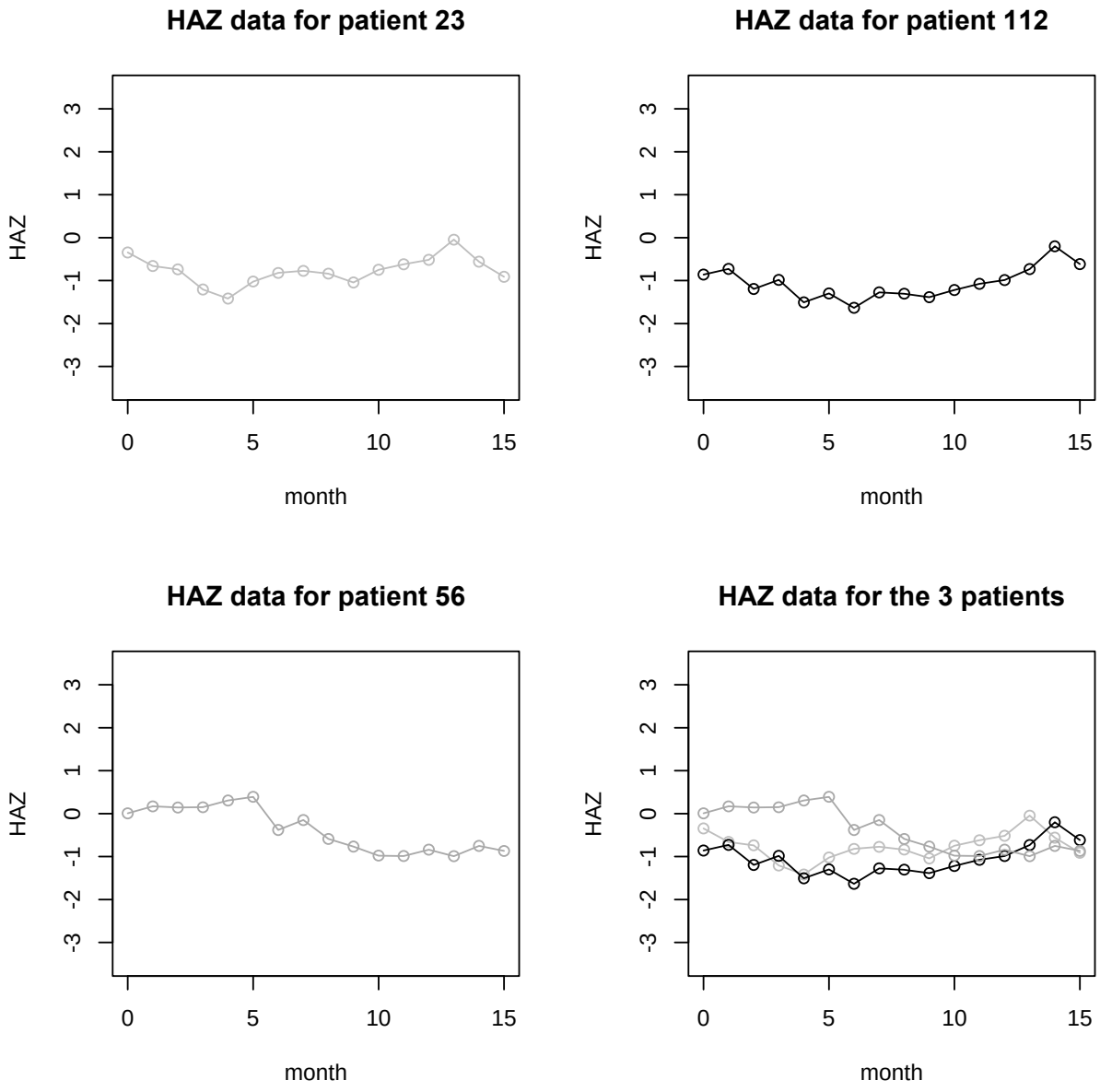


Figure 2: The graphs depicts the fragment one HAZ data for patients 23, 56, and 112 as well as a graphical comparison of the three subjects.

Chapter 4. Methods for Imputation

The datasets we analyze display missing data. The missing data we observe occurs within each curve. For example, a child's HAZ curve may have some missing HAZ value for month 3. For such cases we study the imputation of HAZ data at month 3. There are several methods of data imputation in the literature of longitudinal and functional data analysis. Prior to applying dynamic functional models we require methods of imputation that will empirically provide calculated values to fill these unobserved data. We use seven different methods of imputation that are discussed below. The analysis would also enable us to compare which imputation method to use for missing historic data that would provide a better prediction accuracy for the future HAZ data we study.

There are two different categories of imputation methods which we study, curve-by-curve and pooled curves methods. Pooled curves methods impute for all 197 subjects at the same time. Conversely, curve-by-curve methods do the data imputation at the subject level, that is subject by subject. Methods based on pooled curves might perform better, as they make use of more available data information when doing the imputations. The methods for data imputation we implement are: a) Subject Specific Regression, b) Linear Interpolation and Extrapolation, c) Last Observation Carried Forward, d) Linear Mixed Effects Model, e) Functional Principal Components Analysis, f) Penalized Smoothing with Fourier Bases, and g) Penalized smoothing with Basis functions. Missing data only occurs for fragment one data; that is, HAZ data $Y_{i,l}$ for months l =0, 1, 2, … , 15, for each subject i, i=1,2,…,197. Table 1 contains brief descriptions of each method. We also present each method in detail.

Method	Description
Subject Specific Regression (SSR)	We obtain the predicted value from the model of smooth spline regression.
Linear Inter and Extrapolation (LIE)	Uses the surrounding two data points to approximate a value to fill the gap.
Last Observation Carried Forward (LOCF)	Carries the value of the previous point to fill the gap.
Linear Mixed Effects Model (LME)	Models the between and within observed responses.
Functional Principal Components Analysis (fPCA)	Uses an estimate of the covariance for calculating the prediction of the value for the missing point.
Penalized Smoothing Method I, II (PLS1, PLS2)	Two methods which use basis function expansions (splines for PLS1 and Fourier for PLS2) for each curve and penalized least squares functional regression.

Table 1: This table displays an explanation for each of the seven data imputation methods we use for our study. Curve-by-curve methods can be seen highlighted in light gray, while pooled curves methods are highlighted in dark gray. Each method is also presented separately in the next sections where more details are included.

4.1 Subject Specific Regression

Subject Specific Regression (SSR) obtains the predicted value from the model fit of nonlinear regression with multiple scalar covariates as discussed in Green and Silverman (1994). In this example, we use each month l from l=0,1, …, 15 as our scalar covariates. This method is applied for each curve separately. For each individual subject, there exists a regression equation Y_l =f(l) where 0<l<15 represents our time variable.

To implement this method, we used the `smooth.spline` function in R to fit a cubic smoothing spline to the data (Green and Silverman 1994).

4.2 Linear Inter and Extrapolation

Linear inter and extrapolation (LIE) uses both the data value of the (l-1)th and (l+1)th months to approximate the middle value. To do this, we used the R linear interpolation function `approx`. The `approx` function uses the surrounding two data points to return an interpolated value for the missing point. The function has the ability to return a constant or a function interpolation. We implemented a linear function interpolation approach.

4.3 Last Observation Carried Forward

Last Observation Carried Forward (LOCF) uses the HAZ data vale of the (l-1)th month to impute the missing value at month l. In other words, it carries the value of the previous data point to fill the gap: $Y_{i,l} = Y_{i,l-1}$. By implementing this method of imputation, researches are able to retain the number of subjects, eliminate missingness, and produce a complete data matrix (Overall et al. 2009). A fault in this method of imputation is that it requires that unrealistic assumptions must sometimes be made. This imputation method is often used to compensate for patient drop offs in clinical trials (Overall et al. 2009, Verbeke et al. 2000, Diggle et al. 2002, Hedeker et al. 2006). In this setting, LOCF is used to retain the subject data despite the missingness that comes from the incomplete trial.

4.4 Linear Mixed Effects Model

The linear mixed effects (LME) model is used for modelling longitudinal data (Cnaan et al. 1997, Zhang et al. 2001, Verbeke et al. 2009). LME models the between and within observed responses. This data modeling allows for non-independence and within-subject clustering (Grajeda et al. 2016). The work in Grajeda et al. 2016 proposed the use of LME to approximate a functional line using truncated polynomial splines (cubic) in the

context of child growth data. The splines are comprised of basis functions. The number of basis functions used to approximate the line are based upon the number of knots, or anchor points, used. In our numerical study, we use 3 knots located at months 3, 7, and 11. Based on the anchor points, the basis functions come together to form our approximated function. LME performs these approximations for curves for all 197 subjects in one step for model fitting. Imputation is based on the resulting model fit.

To implement this method of imputation, we used the `lme` function in R. This function is used to fit a linear mixed-effects model.

4.5 Functional Principal Components Analysis

Functional Principal Components Analysis is a very popular method in the functional data analysis arena. It has been used in different contexts, but in our context we are using it as an imputation method. Functional Principal Components Analysis (fPCA) uses an estimate of the variance and covariance functions for generating the prediction of the missing point. The idea of fPCA is that the entire data variability can be reduced to approximately three components. We used the fPCA method from Goldsmith et al. (2013) and the `fpca.sc` R function (Goldsmith et al. 2019). Chapter 3 contains a discussion about fPCA.

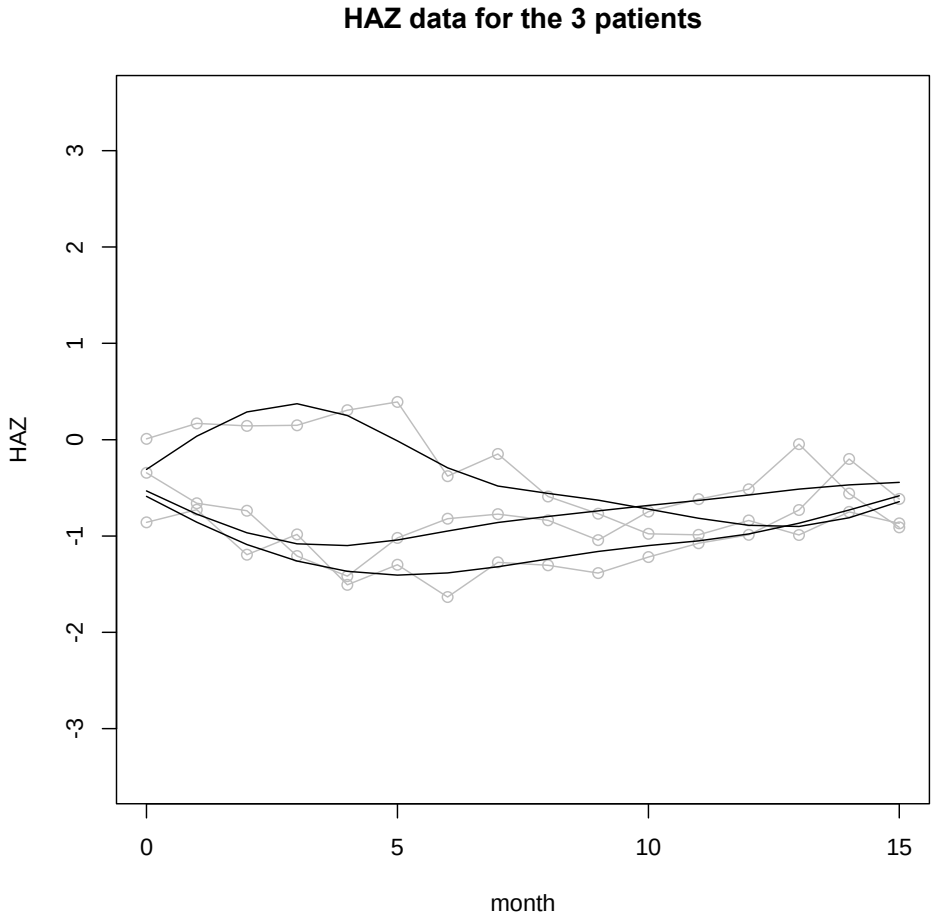


Figure 3: This graph depicts the functional principal components fit for subjects 23, 56, and 112. The gray points show the observed values for the given subjects. The solid black lines display the fPCA fit. This is based upon the availability of the complete fragment 1 data set.

In our experiment, we found that the three components explain 90.14%, 6.80%, and 3.04% of the variability respectively when using the complete data for fragment 1 for months 0-15. Using this method, an estimated fPCA line is fitted for each subject for each of the data scenarios considered, and a value is imputed from the estimated line to compensate for the missingness.

We applied fPCA as an imputation method. In our study data context where missingness was observed, we generated the smoothing trajectories using the available data

points given that one or more points have been removed. This differs from the method we see in the graph of Figure 3 as we had a complete fragment one available.

4.6 Penalized Smoothing Method 1

Penalized smoothing (PLS) is a method used on data whose observed values show a substantial amount of noise. This noise causes the functional objects related to the basis functions to inherit variability and appear "wiggly". To combat these road blocks, penalized smoothing typically uses a large number of basis functions. There is a flexibility about the type of basis used. Besides splines, a popular type is Fourier. Fourier basis functions assume that the underlying function of the data is periodic. We use penalized smoothing that makes use of B-splines and Fourier basis functions. This is a point of variation between this method and LME, which we discussed earlier. Penalized smoothing typically uses a larger number of splines. PLS1 is a method related to basis function expansions. The term $Y_{i,l}$ is expressed using b-spline basis functions.

It is important to be able to express functional data using basis expansion. This method of basis expansion is depicted below where Φ_m are the collection of basis functions i.e (splines, wavelets, sine/cosine functions). These are evaluated at some grid of points $l \in [0,15]$ the same grid for all curves $Y_{i,l}$.

$$Y_{i,l} \approx \sum_{m=1}^{M} c_{i,m}\Phi_m(l), \;\; 1 \leq i \leq n \qquad (5)$$

This expansion assumes that the data can be approximated as a linear combination of M basic shapes Φ_m. M is smaller than the number of observed points per curve. This puts the curves on a common domain given by the basis functions, making them more

readily comparable. Each curve $Y_{i,l}$ is associated with the model coefficients $c_{i,m}$, m=1,2,…,M.

The goal is to find the values of the coefficients which will minimize the penalized sum of squares (PSS_λ(c1,c2….cm)). In summary, this method uses basis function expansion for each curve and regularized regression. This method employs a penalized least squares functional regression approach to estimation:

$$PSS_\lambda(c_1,c_2\ldots.c_m)=\sum_{i,l}(Y_{i,l}-\hat{Y}_{i,l})^2+\lambda\times PEN_2(Y) \quad (6)$$

where λ is the value for the penalty and $PEN_2(Y)=\int[D^2Y_l]^2dl$. The penalty is related to the integrated squared second derivative (Ramsay and Silverman 2005, Ch. 5). In method PLS1, the basis functions $\Phi_m(l)$ employed are B-splines. In our study, we use 10 basis functions, where $\lambda=0.09$. We employed the R function `Data2fd` from the `fda` library (Ramsay et al. 2019). We selected our lambda value based on related implementations where λ values ranged from 0 to 1 (Ramsay et al. 2019).

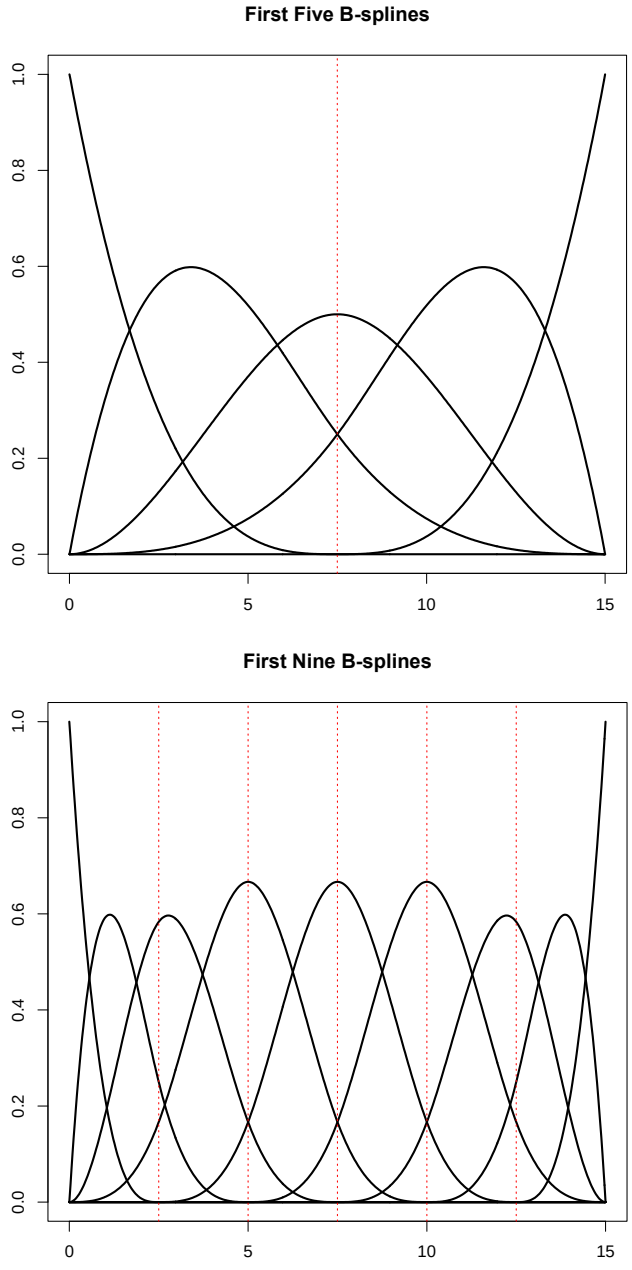
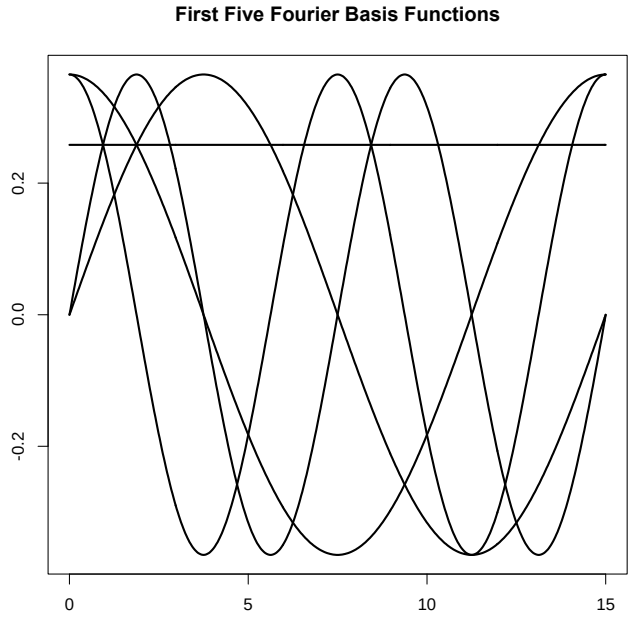


Figure 4: The above graphs are two examples of B- splines similar to the ones we use in our first method of penalized smoothing PLS1.

4.7 Penalized Smoothing Method 2

The difference between this and PLS1 is that the basis functions $\Phi_m(l)$ employed are Fourier.



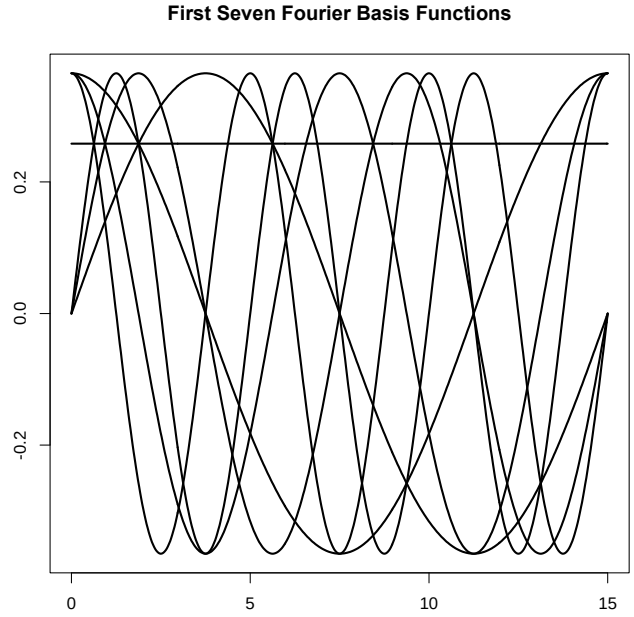


Figure 5: The above graphs are two examples of Fourier basis functions similar to the ones we use in our second method of penalized smoothing.

Chapter 5: Methods of Dynamic Prediction

In order to obtain the full prediction of the future HAZ trajectory, there are several different methods of dynamic prediction that can be applied. Here we focus on functional models for dynamic prediction (Ivanescu et al. 2017). We describe these methods as they relate to the prediction methods used for this project. We employ dynamic methods of prediction due to the complexities of our data set. Some recent applications of dynamic prediction include: using call center data from the beginning of the day to predict call volume for the end of the day (Goldberg et al. 2014), using length-for-age, weight-for-length, and weight-for-age z scores to identify children at risk of delayed growth (Leroux et al. 2018), and using real time traffic flow data to predict an up-to-date traffic flow trajectory (Chiou 2012). It is considered dynamic because the unobserved curve data for future months is being predicted from data that occurred in this history of the curve. It is our interest to look dynamically at our historic data to predict and analyze our future data.

We examine a number of methods of Dynamic Prediction including the Benchmark Dynamic (BENDY) method, the Dynamic Linear Model, Dynamic Penalized Functional Regression, and Dynamic Penalized Function-on-Function Regression (Ivanescu et al. 2017). Chapter 7 contains some numerical analyses that compares the performance of these methods for dynamic prediction. In Chapter 6 we apply these methods to a simulation study. The observed historic HAZ data is denoted by $Y_{i,l}$ for months l=0,…,15 for fragment 1 data. For fragment 2 data, the data is denoted by $Y_{i,\tilde{t}}$ for months $\tilde{t}$ = 30,…,45. We will refer to fragment 1 as historic data, and fragment 2 as future data. For example, as described in Table 2, HAZ data for child i at month 0 will be depicted as $Y_{i,0}$ and HAZ data for child

i at month 15 will be depicted as $Y_{i,15}$. Similarly, for fragment 2, $Y_{i,30}$ represents HAZ data for child i at month 30.

Fragment 1 data (months 0-15)	**Fragment 2 data (months 30-45)**
$\boldsymbol{Y_{i,l}}$	$\boldsymbol{Y_{i,\tilde{t}}}$
HAZ value at month l for child i, where l=0,…,15 and i=1,2,…, 197.	***HAZ value at month $\tilde{t}$ for child i, where*** $\tilde{t}$ = 30,…,45 and i=1,2,…, 197.

Table 2: This table depicts the notation that will be used throughout this paper to denote subjects across different times within fragment one and fragment two.

5.1 Benchmark Dynamic Method (BENDY)

The first dynamic model we use is the BENDY model. The Benchmark Dynamic (BENDY) method uses the first and last point in the historic data set. The historic data is then used to predict the data point for each future month $\tilde{t}$, where there is some distance between the historic and future data. We use notation $Y_{i,0}$ to reflect the HAZ value for subject i at the start of fragment one, that is at month 0, and $Y_{i,15}$ to reflect the HAZ value within fragment one at month 15. For future HAZ data we denote by $Y_{i,\tilde{t}}$ the HAZ value within fragment 2 at month $\tilde{t}$. For our purpose $Y_{i,l} = HAZ_{i,l}$, but this provides the generic BENDY model notation

$$Y_{i,\tilde{t}} = Y_{i,0}\beta_{0,15,\tilde{t}} + Y_{i,15}\beta_{15,15,\tilde{t}} + \epsilon_{i,\tilde{t}}. \quad (7)$$

The BENDY model listed above is based upon prediction for one future data point for HAZ, specifically at month $\tilde{t}$. In our case, we are using historic HAZ data to predict the

HAZ score for future months. In our context $Y_{i,\tilde{t}}$ denotes $HAZ_{i,\tilde{t}}$, $Y_{i,0}$ denotes $HAZ_{i,0}$, and $Y_{i,15}$ denotes $HAZ_{i,15}$.

If we were to consider a secondary historic data source for prediction, our model would be adjusted to include $Z_{i,0}$ and $Z_{i,15}$.

$$Y_{i,\tilde{t}} = Y_{i,0}\beta_{0,15,\tilde{t}} + Y_{i,15}\beta_{15,15,\tilde{t}} + Z_{i,0}\gamma_{0,15,\tilde{t}} + Z_{i,15}\gamma_{15,15,\tilde{t}} + \epsilon_{i,\tilde{t}} \quad (8)$$

An example of a potential historic data source Z we could consider would be a Weight for Age Z-score (WAZ) data set. In this instance, we would be using a combination of HAZ and WAZ data to predict the future HAZ data. Model parameters are indexed based on the time coordinates of the historic data and future data where prediction is done.

5.2 Dynamic Linear Model (DLM)

The Dynamic Linear Model uses historic data to predict future scalar responses similar to the BENDY Model (Ivanescu et al. 2017). The primary difference between the two is that the DLM uses all available data points within the historic data in addition to the first and last historic data points. Below is the DLM model that further expands on the BENDY model to the context of our historic data setting. This is reflected in the formula for the DLM model

$$Y_{i,\tilde{t}} = \sum_{l=0}^{15} Y_{i,l}\beta_{l,15,\tilde{t}} + \epsilon_{i,\tilde{t}}. \quad (9)$$

DPFR and DPFFR are some other models that make use of the entire historical data set from months 0 to l5. The DLM model takes the sum of the covariates while the DPFR and DPFFR integrate the functional covariates across all the months in fragment 1.

5.3 Dynamic Penalized Function Regression (DPFR)

Dynamic Penalized Function Regression is a modified version of DLM in that it imposes penalized smoothing onto the model coefficients. Dynamic Penalized Function Regression (DPFR) is represented by the same model as DPFFR which will be discussed in the following section. The model is depicted below

$$Y_{i,\tilde{t}} = \int_{l=0}^{15} Y_{i,l}\beta_{l,15,\tilde{t}}\, dl + \epsilon_{i,\tilde{t}}. \tag{10}$$

A difference between the two is that in DPFR, the response variable $Y_{i,\tilde{t}}$ is considered as a scalar response, not a functional response like in DPFFR. This is due to differences in the smoothing components of these two models. The penalized smoothing in DPFR consists of smoothing for the regression coefficient in the direction of only historical data whereas, DPFFR is able to perform smoothing for the functional regression model coefficients in multiple dimensions.

5.4 Dynamic Penalized Function-on-Function Regression (DPFFR)

Throughout this paper's numerical study in Chapter 6, we employ Dynamic Penalized Function-on-Function Regression (DPFFR) as a method of predicting the functional response $Y_{i,\tilde{t}}$ of subject i over fragment two. This differs from the methods of BENDY, DLM, and DPFR in that is predicts functional responses rather than targeting scalar outcomes. We use historic data to obtain predictions for the estimates of future values $Y_{i,\tilde{t}}$. The model for DPFFR is

$$Y_{i,\tilde{t}} = \int_{l=0}^{15} Y_{i,l}\beta_{l,15,\tilde{t}}\, dl + \epsilon_{i,\tilde{t}}. \tag{11}$$

When reviewing the DLM, DPFR, and DPFFR models of prediction, it is presumed that the DPFFR model will provide the most accurate predictions, while DPFR and DLM might follow behind respectively.

All methods of prediction will employ the technique of leave one-curve out cross-validation (Ivanescu et al. 2017), forming a test set of 1 subject, and a training set of 196. We will use this technique in order to align with the method of model prediction validation (Ivanescu et al. 2016).

Chapter 6. Numerical Study

Our empirical study features a simulation of fully observed (dense) data for fragment 2 (HAZ data for months 30-45). Data for months 0-15 was available (Ivanescu et al. 2017) for fragment 1 data. For this simulation we used the data for fragment 1 (HAZ data for months 0-15) as input data in the data generation model. We simulated 100 different datasets for HAZ at months 30-45. Figure 6 depicts an example of simulated dataset where fragment 2 data was simulated using fragment 1 data.

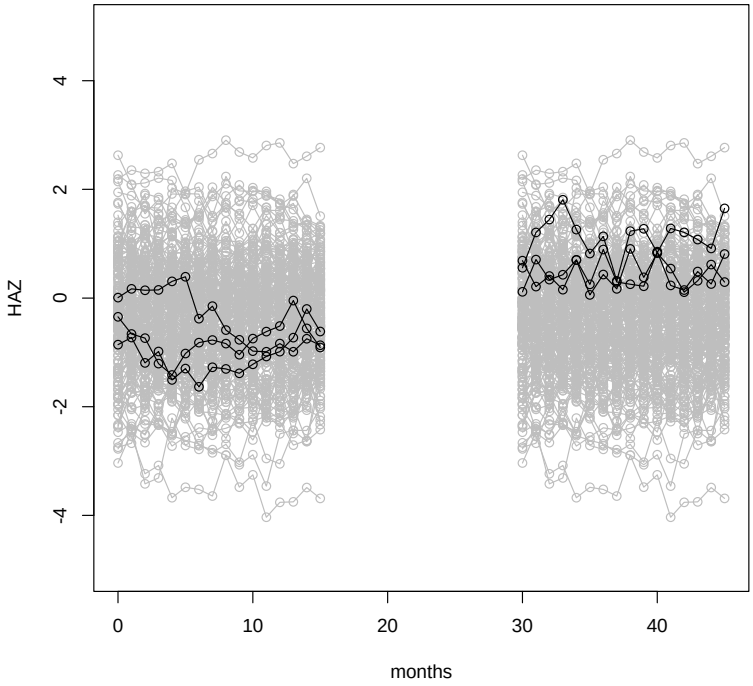


Figure 6: The above graph depicts the fragment 1 and fragment 2 data for all 197 subjects. The model used was the DPFFR model. Highlighted are subjects 23, 56, and 112.

Simulations make use of the DPFFR statistical model (Ivanescu et al. 2017) that was described in Section 5.4. The functional parameters chosen are similar to the work in Goldsmith et al. (2012) and Ivanescu et al. (2015). Specifically, we employed a functional

bivariate slope of the form $\beta_{l,15,\tilde{t}} = \frac{1}{2.5}\left(\cos(\frac{2\pi t}{16}) \times \sqrt{\frac{\tilde{t}}{\tilde{t}^{.4}}}\right)$. We represent our DPFFR model as $Y_{i,\tilde{t}} = \zeta(\tilde{t}) + \int_{l=0}^{15} Y_{i,l}\beta_{l,15,\tilde{t}}dl + \varepsilon_{i,\tilde{t}}$, where our errors were assumed to be normally distributed as $\varepsilon_{i,\tilde{t}} \sim N(0, 0.10^2)$ and $\zeta(\tilde{t}) = e^{-10(\tilde{t}-34.5)^2}$ was the intercept function. There were 100 simulated datasets for fragment 2 data. The bivariate model parameter is displayed in Figure 7. Other values for the composition of the $\beta_{l,15,\tilde{t}}$ parameter would yield different datasets $Y_{i,\tilde{t}}$.

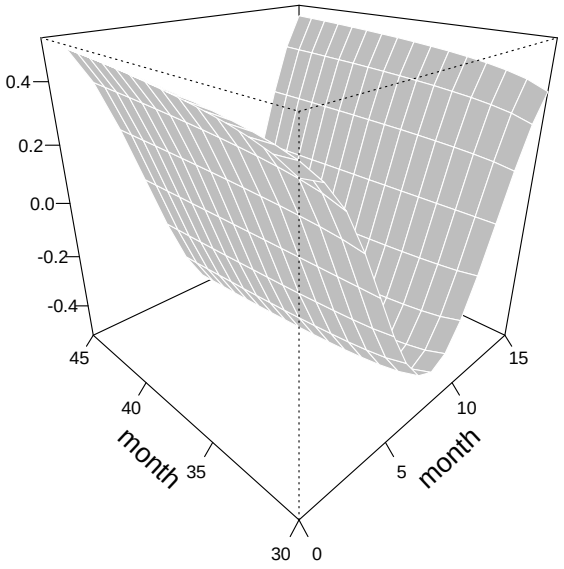


Figure 7: Illustrated above is the beta function employed in the DPFFR model for data generation.

6.1 Data Generation

Our initial HAZ data set provides the height-for-age Z scores (HAZ) for 197 patients across months 0-15 (fragment 1). We will use the data from months (0-15) in order to predict the data in months 30-45 (fragment 2). We examine the use of a dynamic prediction model.

We use the DPFFR (Dynamic Penalized Function-on-Function Regression) model of dynamic prediction. We employ the technique of leave one curve-out cross validation (Ivanescu et al 2017) in order to obtain prediction for each curve i at month $\tilde{t}$: $\hat{Y}_{i,\tilde{t}}$.

6.2 Simulation of Missing Data

For months 1-14 we simulate missingness. We only simulate missing data for one month at a time. Table 1 Column I1 contains MSEP metric for prediction of $Y_{i,\tilde{t}}$ when month 1 was simulated as missing and then imputed using imputation methods employed, such as LOCF.

In the numerical study we simulate missing values for data at months 1-14. Data from only one month is removed in each iteration. Data for the first and last month (month 0 and month 15) are considered observed. To begin with, we remove HAZ data at month 1 for all subjects. Imputation methods can provide a plug in or calculated value for month 1 data for all subjects. Then, dynamic prediction model DPFFR is employed to obtain predicted HAZ data at months 30-45 using data from months 0-15 as predictive information. Overall, we consider removal of each month separately in the history of HAZ data, starting with month 1 through month 14, where only one month is removed for all subjects each time. In doing this, we will examine which month removal provides a prediction performance that suggests that the missing HAZ data and subsequent imputation will have the least effect on the overall dynamic prediction. The goal is to suggest four months where HAZ data points are planned for removal from the months 0-15 HAZ trajectory. After studying the removal of one month, we compare the prediction performance across months to decide on the candidate month for removal. In the next step, the strategy we follow is to remove the second HAZ data for a given HAZ curve given the

first point has already been removed. This will keep the first missing data uniform amongst all subjects, and the second missing data will be systematically generated to be the same among all subjects.

6.3 Metrics

Several metrics were used in this study to compare the different methods applied. We used the Mean Squared Error for Prediction (MSEP), Mean Squared Error for Estimation (MSEE), and Akaike Information Criterion (AIC).

Mean Squared Error: MSEP is calculated for each dataset as the average of the squared of errors between the actual value Y and the predicted value, $\hat{Y}$. The metric is used to track how great of an impact the data point removal and imputation has on our predicted values for moths 30-45. An equation representing this function for a given dataset is:

$$\text{MSE}= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{15}\sum_{\tilde{t}=30}^{45}\left(Y_{i,\tilde{t}} - \hat{Y}_{i,\tilde{t}}\right)^2 \quad (12)$$

where $\hat{Y}$ represents our predicted value and Y represents the value for the response Y at month $\tilde{t}$ in fragment 2. We calculate the MSEP for each dataset and we report MSEP as the average across 100 datasets. MSE of Estimation (MSEE) is used as metric for estimation of β. AIC is used to estimate the quality of a set of statistical models relative to one another for a given set of data. We provide AIC metrics (Wood 2006) derived from R implementations corresponding to the functional models (Ivanescu et al. 2017) for dynamic prediction.

6.4 Results for Dynamic Prediction

We compare several data imputation mechanisms for longitudinal functional data. Using the data generated in the numerical study, imputation is performed for the seven

imputation methods SSR, LIE, LOCF, LME, fPCA, PLS1 and PLS2. These techniques were described in Chapter 4.

Imputation Step 1

Imputation Step 1 contains an analysis for the study of imputation for one month among all historic HAZ data (months 0-15). Because we only simulate missingness for the inner data points, months 1-14 are considered candidates for removal of HAZ data and subsequent imputation.

Below is a table tracking the Mean Squared Error (MSEP) for DPFFR prediction after removing each of the 14 inner data points. Table 1 Column Im1 (Imputation for month 1) contains MSEP for DPFFR prediction for future HAZ $Y_{i,\tilde{t}}$ when historic month 1 HAZ data was simulated as missing and then imputed using methods SSR, LIE, LOCF, LME, fPCA, PLS1 and PLS2. Columns Im2 (Imputation for month 2) through Im14 (Imputation for month 14) contain similar quantities when HAZ at each of months 2-14 was considered as a candidate for removal. It is our goal to repeat this process until we find several months to consider for removal from the historic HAZ data from months 1-14. The step of dynamic prediction is done after the imputation of historic data. We employ DPFFR for dynamic prediction of HAZ data at months 30-45. This study considers data generated from the numerical study where we simulated 100 different HAZ datasets for fragment 2 data (months 30-45). The prediction is then compared with the true value to determine the MSEP metric across months 30-45, across all subjects, and across all 100 simulated HAZ datasets. Numbers in Table 3 contain average MSEP values across 100 simulated datasets. We will also be measuring the procedures in terms of the Akaike Information Criterion (AIC), and the Mean Squared Error for Estimation (MSEE) where we compare our

estimated β with our true β value. Tables in the Appendix show the results for these metrics. After we find the initial least problematic point to remove, we will remove another point given the first is already out of the sampling design. We will create

another set of tables for each point removed to develop the sampling schedule.

Summarized Results for Imputation Step 1

The Mean Squared Error for Prediction (MSEP) ranged from 10.08 to 12.97, where 10.08 was an optimal prediction that considered all available data fully observed. Considering that a smaller MSEP is desired, months 4 and 12 are shown to be candidate months for HAZ data removal as their imputation yields the smallest MSEP across all months 1-14 considered as candidates for removal. Columns Im4 and Im12 had the smallest MSEP among all columns Im1-Im14. These results are confirmed by further analysis of the additional metrics.

The imputation methods that appear to be the most effective overall for approximation of the missing HAZ data value are seen to be Functional Principal Component Analysis (fPCA) and the Linear Mixed Effect Model (LME), as these exhibit smaller MSEP values. These are both pooled curves methods, meaning that for their implementation the entire set of curves are needed for application of the data imputation method.

	SSR	LIE	LOCF	LME	fPCA	PLS1	PLS2
Im1	12.5	11.97	11.83	11.81	11.82	11.88	11.99
Im2	11.52	11.46	11.89	11.3	11.32	11.38	11.42
Im3	10.66	10.57	10.63	10.53	10.54	10.55	10.57
Im4	10.08	10.08	10.08	10.08	10.08	10.08	10.08
Im5	10.46	10.43	10.53	10.36	10.37	10.41	10.44
Im6	11.58	11.42	11.72	11.17	11.17	11.3	11.38
Im7	12.02	11.88	12.35	11.73	11.73	11.8	11.86
Im8	12.35	12.39	12.97	11.91	11.9	12.1	12.23
Im9	11.86	11.86	12.33	11.57	11.56	11.71	11.8
Im10	11.13	11.15	11.54	10.85	10.84	11.01	11.13
Im11	10.32	10.33	10.45	10.31	10.31	10.32	10.33
Im12	10.08	10.08	10.08	10.08	10.08	10.08	10.08
Im13	10.4	10.37	10.44	10.36	10.36	10.36	10.37
Im14	11.19	11.17	11.54	11.11	11.11	11.12	11.17

Table 3. Results for MSEP for the case of one-month data that is missing. Results are displayed as MSEP x 100.

The Mean Squared Error for Estimation (MSEE) results shown in the tables in the Appendix range from .0215 to .5218, where .0215 is an optimal estimation of β model parameter. Given this metric of estimation we conclude that months 4 and 12 are indicated as candidates for removal because MSEE is smallest for columns Im4 and Im12. The imputation method that is seen to be the most effective for estimation of β is Functional

Principal Component Analysis because the row labeled fPCA contains the smallest MSEE across all imputation methods.

The Akaike Information Criterion (AIC) results posted in the Appendix range from .1704 to .2460, where .1704 is an optimal value for this metric. Based on this metric, months 4 and 12 are rated equally, because AIC was smallest among all results in the table. Based on this metric alone, it appears that either one could be removed relatively unproblematically.

Based upon the results highlighted by the above metrics, month 4 is unanimously the least problematic month to remove when studying the DPPFR model. It appears that its removal would have negligible effect if considered missing. Although, is also apparent that month 12 is very similarly unproblematic. In terms of imputation method, fPCA is the most effective across the board. We have determined that we will remove both months 4 and months 12 before going on to further this sampling study design. When looking at the least problematic months to remove, it appears our results were cohesive regardless of imputation method. To continue, we ran the next step of the sampling design study given that HAZ data at months 4 and 12 has already been removed.

Imputation Step 2

Imputation Step 2 contains an analysis for the study of imputation for one month among historic HAZ data, given that HAZ data at months 4 and 12 were assumed removed from the HAZ historic data sampling design. Months 1-3, 5-11, 13-14 are considered candidates for removal of HAZ data and subsequent imputation. Table 4 contains MSEP results for imputation step 2 and additional metrics were placed in the Appendix.

Summarized Results for Imputation Step 2

The MSEP results ranged from 10.08 to 13.10. A value of 10.08 was calculated to be the MSEP corresponding to a fully observed historic HAZ dataset. This is a slightly larger range than that of the previous simulation. Based upon the results, it appears month 11 and 13 are the least problematic to remove. It is also important to note that months 3 and 5 appear relatively unproblematic as well. Recall that the above-mentioned months border the months which we removed in the previous step.

	SSR	LIE	LOFC	LME	fPCA	PLS1	PLS2
Im1	12.5	12.00	11.94	11.89	11.89	11.91	11.99
Im2	11.6	11.44	11.91	11.34	11.34	11.39	11.43
Im3	10.74	10.58	10.63	10.55	10.56	10.57	10.59
Im4							
Im5	10.57	10.50	10.58	10.39	10.39	10.47	10.54
Im6	11.74	11.42	11.66	11.19	11.2	11.34	11.42
Im7	12.11	11.92	12.39	11.77	11.77	11.81	11.86
Im8	12.44	12.45	13.1	11.96	11.97	12.14	12.29
Im9	11.99	11.90	12.45	11.63	11.64	11.74	11.83
Im10	11.22	11.12	11.53	10.86	10.85	11.01	11.12
Im11	10.41	10.36	10.42	10.32	10.32	10.34	10.35
Im12							
Im13	10.44	10.38	10.45	10.37	10.37	10.37	10.37
Im14	11.32	11.21	11.43	11.18	11.18	11.19	11.25

Table 4. Results for MSEP for Imputation Step 2. Results are displayed as MSEP x 100.

The imputation methods that appear to be the most effective for DPFFR prediction performance are seen to be Functional Principal Component Analysis and the Linear Mixed Effect Model.

The Mean Squared Error for Estimation (MSEE) results posted in the Appendix range from .023 to .697. Similar to MSEP, this range is slightly wider than that from the previous step. After examining this table, it is confirmed that months 3, 5, 11, and 13 are least problematic to remove. The imputation method that is seen to be the most effective for estimation results of β model parameter is Functional Principal Component Analysis.

The Akaike Information Criterion (AIC) results posted in the Appendix ranged from .170 to .249. Based on this metric, month 11 is a candidate and month 13 falls close behind.

Based upon the results highlighted by the above metrics, months 11 and 13 are equally unproblematic to remove. The removal of these points will have the least impact on MSEP, MSEE, and model accuracy. We have determined that we could remove both months 11 and 13 at the conclusion of the Imputation Step 2. Our results are cohesive amongst all the metrics, but we also think it is important to note the results of months 3 and 5 as well. They were slightly more problematic to remove, but at this stage, we do not have sufficient evidence to suggest anything else for removal in addition to months 11 and 13.

Figure 8 depicts the conclusion of our sampling study. The future curves are suggested to be predicted using the remaining historic data points illustrated in black while acknowledging a corresponding change in prediction performance.

Upon successfully suggesting the removal of HAZ data at months 4, 11, 12, and 13 based on our numerical study conducted in this chapter we completed our goals for our sampling design study.

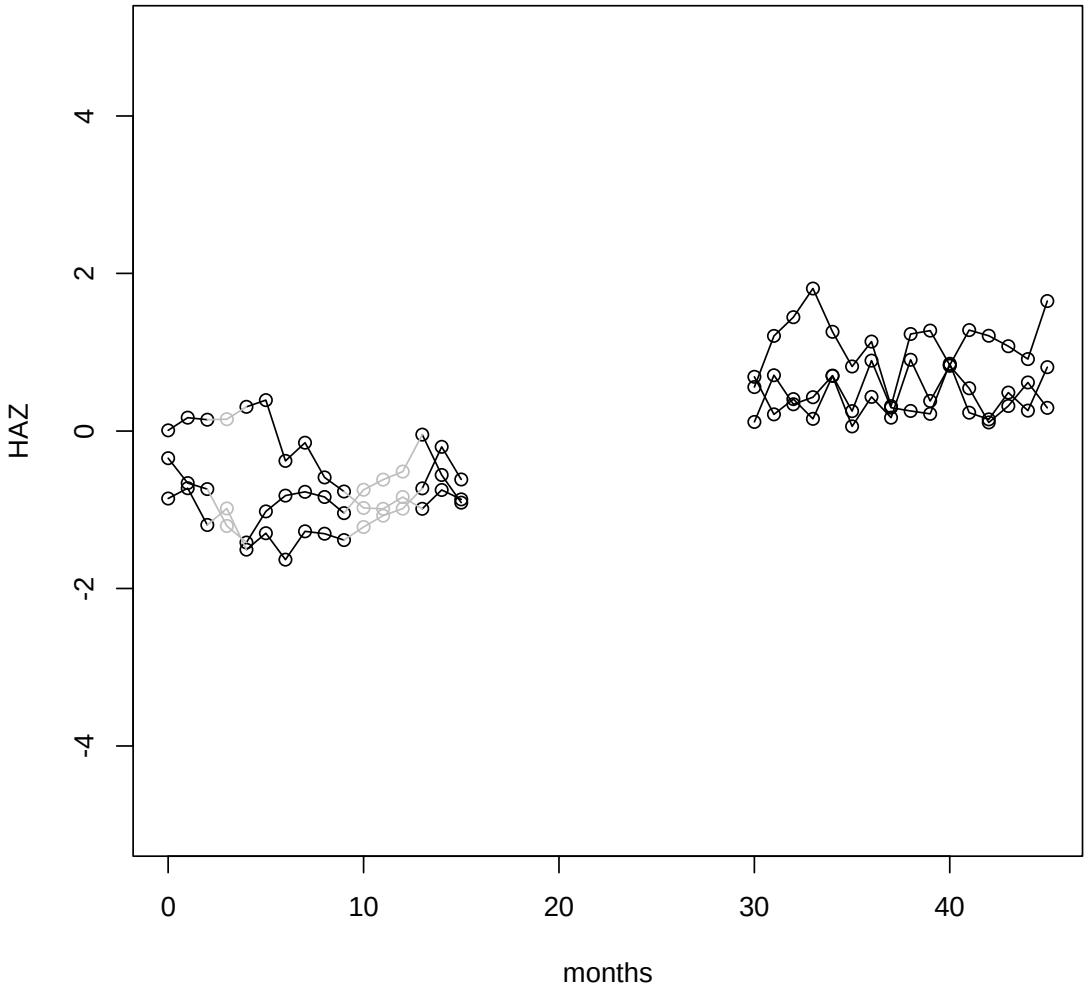


Figure 8: Illustrated above is the HAZ data for 3 subjects post removal of months 4, 11, 12, and 13. The gray points show the location of suggested HAZ data that are suggested for removal in the sampling study.

Chapter 7. Dynamic Prediction for Longitudinal Functional Data

In this chapter we conduct an additional numerical study for comparing several dynamic prediction methods in the context of longitudinal functional data.

7.1 HAZ Data

We now examine the historic HAZ data set (months 0-15) as two distinct fragments. Recall that we have 15 months of data for 197 subjects. We intend to further study additional methods of dynamic prediction for longitudinal functional data using the different fragments of this longitudinal data. We take months 0 to 5 as our historic data. It is our intent to measure the accuracy of nine different methods of dynamic prediction by using historic HAZ data at months 0 to 5 to predict HAZ data at future months 10 to 15. We will use the notation $Y_{i,l}$ for months l=0,..,5 (historic data). To denote future HAZ data, we will use $Y_{i,\tilde{t}}$ for months $\tilde{t}$ = 10,..,15.

7.2 Methods of Dynamic Prediction

We investigate several dynamic prediction methods presented in Chapter 5 (BENDY, DLM, DPFR, DPFFR) for our HAZ data setting in this chapter. The BENDY, DLM, DPFR, and DPFFR dynamic models have been discussed earlier and use the method of leave one-curve out CV (Ivanescu et al. 2017). However, these methods have not been investigated in the context of longitudinal functional data. Several additional methods based on Nearest Neighbors (NN) denoted by NN1, NN2, NN3, NN4, NN5 are methods based on the principles of nearest neighbors and were applied here for dynamic prediction of longitudinal functional data. We discuss these methods next.

NN1: This method of dynamic prediction calculates n-1 distances between curves where n is equal to 197. Given historic data for a given subject, 196 distances are computed for historic data. Each distance is computed between the HAZ curve for a given subject i and all the remaining curves consisting of n-1 curves. Each subject serves as subject i for the calculation of these distances. This procedure is similar to the method of leave one-curve out validation. The purpose of this method is to determine which subject behaves the closest to the subject we aim to predict. When the subject within the closest distance (smallest distance) is identified, we use that subject's future HAZ data as an estimate for the points we wish to predict.

NN2: This method is similar to NN1, but it incorporates the 2 subjects that behave the closest to the subject we aim to predict. It uses a pointwise average to calculate the estimation for the points we wish to predict. We examine several similar methods where we incorporate the 2^{nd}, 3^{rd}, 4^{th}, and 5^{th} closest subjects for estimations. We will refer to these methods as NN2, **NN3**, **NN4**, and **NN5** respectively.

7.3 Results

Results are displayed in Table 5. Using mean squared error of prediction (MSEP) as the metric, we discovered that DPFFR is the most efficient method and has the smallest MSEP. This method is closely followed by DLM and DPFR. Amongst the nearest neighbors methods, it can be seen that prediction accuracy increases as additional subjects are incorporated in the calculations.

Method of Dynamic Prediction	Mean Squared Error of Prediction (MSEP)
BENDY	22.799
DLM	19.253
DPFR	19.296
DPFFR	19.066
NN1	45.802
NN2	33.542
NN3	30.965
NN4	28.865
NN5	27.164

Table 5: The table above reflects the Mean Squared Error for Prediction (MSEP) values across our 9 methods of dynamic prediction. Results are displayed as MSEP x 100.

Chapter 8: Conclusion

The work proposed considered the data setting of longitudinal functional data. We discussed a setting where sampling studies can be designed when using dynamic models for prediction. The methods deal with data imputation methods. The imputation methods that use pooled curves performed better than imputation methods performed only at the curve level. Implementations used the method of dynamic prediction DPFFR (dynamic function-on-function regression). There are several other methods of dynamic prediction that can be studied in the context of longitudinal functional data, such as BENDY, DLM, DPFR, or other methods based on nearest neighbors.

References

Bunea, F., Ivanescu, A. E., and Wegkamp, M. H. (2011). Adaptive inference for the mean of a Gaussian process in functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 531-558.

Chén, O. Y., Xiao, L., Caffo, B. S., Lindquist, M. A., Schrack, J. A., Ferrucci, L., and Crainiceanu, C. M. (2016). A marginal approach to longitudinal functional data for analyzing daily physical activity patterns. Accessed on March 22, 2019 at http://oliverychen.github.io/files/doc/LFDA.pdf

Chiou, J. M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. The Annals of Applied Statistics, 6(4), 1588-1614.

Cnaan, A., Laird, N. M., and Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Statistics in medicine, 16(20), 2349-2380.

Diggle, P., J., Heagerty, P., Heagerty, P. J., Liang, K. Y., and Zeger, S. (2002). Analysis of longitudinal data. Oxford University Press. Oxford.

Fan, Y., Foutz, N., James, G. and Jank, W. (2014) Functional response additive model estimation with online virtual stock markets. The Annals of Applied Statistics, 8, 2435-2460.

Goldberg, Y., Ritov, Y. A., and Mandelbaum, A. (2014). Predicting the continuation of a function with applications to call center data. Journal of Statistical Planning and Inference, 147, 53-65.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. Journal of the Royal Statistical Society: Series C (Applied Statistics), 61(3), 453-469.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. Biometrics, 69(1), 41-51.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, L., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2019). `refund`: Regression with Functional Data. R package version 0.1-17. https://CRAN.R-project.org/package=refund

Grajeda, L. M., Ivanescu, A. E., Saito, M., Crainiceanu, C. M., Jaganath, D., Gilman, R. H., Crabtree, J. E., Kelleher, D., Cabrera, L., Cama, V., and Checkley, W. (2016). Modelling subject-specific childhood growth using linear mixed-effect models with cubic regression splines. Emerging Themes in Epidemiology, 13, 1-13.

Green, P. J., and Silverman, B. W. (1994). Nonparametric regression and generalized linear models. Chapman & Hall New York.

Hedeker, D. R., and Gibbons, R. D. (2006). Longitudinal data analysis. Hoboken, NJ: Wiley-Interscience.

Ieva, F., and Paganoni, A. M. (2016). Risk prediction for myocardial infarction via generalized functional regression models. Statistical Methods in Medical Research, 25(4), 1648-1660.

Islam, M. N., Staicu, A. M., and van Heugten, E. (2016). Longitudinal dynamic functional regression. Accessed on March 22, 2019 at https://arxiv.org/abs/1611.01831

Ivanescu, A. E., Staicu, A. M., Scheipl, F., and Greven, S. (2015). Penalized function-on-function regression. Computational Statistics, 30(2), 539-568.

Ivanescu, A. E., Li, P., George, B., Brown, A. W., Keith, S. W., Raju, D., and Allison, D. B. (2016). The importance of prediction model validation and assessment in obesity and nutrition research. International Journal of Obesity, 40(6), 887.

Ivanescu, A. E., Crainiceanu, C. M., and Checkley, W. (2017). Dynamic child growth prediction: A comparative methods approach. Statistical Modelling, 17(6), 468-493.

Ji, H., and Müller, H. G. (2017). Optimal designs for longitudinal and functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3), 859-876.

Kokoszka, P., and Reimherr, M. (2017). Introduction to functional data analysis. CRC Press, Boca Raton, FL.

Leroux, A., Xiao, L., Crainiceanu, C., and Checkley, W. (2018). Dynamic prediction in functional concurrent regression with an application to child growth. Statistics in Medicine, 37(8), 1376-1388.

Overall, J. E., Tonidandel, S., and Starbuck, R. R. (2009). Last-observation-carried-forward (LOCF) and tests for difference in mean rates of change in controlled repeated measurements designs with dropouts. Social Science Research, 38(2), 492-503.

Park, S. Y., and Staicu, A-M. (2015). Longitudinal functional data analysis. STAT, 4, 212-226.

Park, S. Y., Xiao, L., Willbur, J. D., Staicu, A. M., and Jumbe, N. N. (2018). A joint design for functional data with application to scheduling ultrasound scans. Computational Statistics and Data Analysis, 122, 101-114.

R Development Core Team (2019), R: A Language and Environment for Statistical Computing, R Core Team. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Ramsay, J. O. and Silverman, B. W. (2005). Functional Data Analysis. 2nd ed. Springer Series in Statistics, Springer, New York.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2019). `fda`: Functional Data Analysis. R package version 2.4.8. https://CRAN.R-project.org/package=fda.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

Shang, H. L. (2017). Forecasting intraday S&P 500 index returns: A functional time series approach. Journal of Forecasting, 36(7), 741-755.

Sørensen, H., Goldsmith, J., and Sangalli, L. M. (2013). An introduction with medical applications to functional data analysis. Statistics in Medicine, 32(30), 5222-5240.

Verbeke, G., and Molenberghs, G. (2009). Linear mixed models for longitudinal data. Springer Science and Business Media. New York.

Wood, S. N. (2006). Generalized Additive Models: An Introduction with R, Chapman & Hall/CRC, New York.

Wrobel, J., Park, S. Y., Staicu, A. M., and Goldsmith, J. (2016). Interactive graphics for functional data analyses. Stat, 5(1), 108-118.

Wu, M., Diez-Roux, A., Raghunathan, T. E., and Sánchez, B. N. (2018). FPCA-based method to select optimal sampling schedules that capture between-subject variability in longitudinal studies. Biometrics, 74(1), 229-238.

Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data. Statistics and Computing, 26(1-2), 409-421.

Yao, F., Müller, H. G., and Wang, J. L. (2005a). Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association, 100(470), 577-590.

Yao, F., Müller, H. G., and Wang, J. L. (2005b). Functional linear regression analysis for longitudinal data. The Annals of Statistics, 33(6), 2873-2903.

Zhang, D., and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics, 57(3), 795-802.

APPENDIX

Included below are several other metrics used in determining the sampling study design for longitudinal functional data setting considered. The metrics presented in this Appendix include AIC and mean squared error for estimation (MSEE) accuracy.

Imputation Step 1:

	SSR	LIE	LOCF	LME	fPCA	PLS1	PLS2
Im1	0.240	0.134	0.554	0.225	0.059	0.135	0.045
Im2	0.121	0.098	0.134	0.130	0.033	0.081	0.056
Im3	0.042	0.049	0.107	0.048	0.023	0.049	0.049
Im4	0.021	0.021	0.020	0.022	0.021	0.021	0.021
Im5	0.048	0.043	0.039	0.055	0.025	0.031	0.032
Im6	0.084	0.059	0.107	0.114	0.023	0.064	0.080
Im7	0.361	0.484	0.671	0.274	0.132	0.444	0.532
Im8	0.117	0.236	0.488	0.111	0.028	0.213	0.294
Im9	0.200	0.271	0.467	0.073	0.086	0.252	0.292
Im10	0.060	0.037	0.034	0.044	0.023	0.041	0.046
Im11	0.043	0.045	0.061	0.026	0.023	0.047	0.052
Im12	0.020	0.020	0.020	0.022	0.022	0.020	0.020
Im13	0.084	0.082	0.103	0.051	0.031	0.074	0.065
Im14	0.425	0.203	0.243	0.109	0.038	0.179	0.069

Table 1: Results for MSEE for Imputation Step 1. Results are displayed as MSEE x 100.

	SSR1	LIE	LOCF	LME	fPCA	PLS1	PLS2
Im1	0.235	0.222	0.218	0.218	0.218	0.220	0.222
Im2	0.210	0.209	0.220	0.205	0.205	0.207	0.208
Im3	0.187	0.185	0.186	0.184	0.184	0.184	0.185
Im4	0.170	0.170	0.170	0.170	0.170	0.170	0.170
Im5	0.182	0.181	0.183	0.179	0.179	0.180	0.181
Im6	0.212	0.208	0.215	0.201	0.201	0.205	0.207
Im7	0.223	0.220	0.231	0.216	0.216	0.218	0.219
Im8	0.231	0.232	0.246	0.220	0.220	0.225	0.228
Im9	0.219	0.219	0.231	0.212	0.212	0.215	0.218
Im10	0.200	0.200	0.211	0.192	0.192	0.197	0.200
Im11	0.177	0.178	0.181	0.177	0.177	0.177	0.178
Im12	0.170	0.170	0.170	0.170	0.170	0.170	0.170
Im13	0.180	0.179	0.181	0.179	0.179	0.179	0.179
Im14	0.202	0.201	0.211	0.199	0.199	0.200	0.201

Table 2: Results for AIC for Imputation Step 1. Results are displayed as AIC /10^4.

Imputation Step 2:

	SSR	LIE	LOCF	LME	fPCA	PLS1	PLS2
Im1	0.257	0.126	0.593	0.241	0.058	0.133	0.057
Im2	0.214	0.126	0.165	0.175	0.036	0.117	0.08
Im3	0.1	0.076	0.114	0.08	0.029	0.075	0.079
Im4							
Im5	0.035	0.031	0.078	0.067	0.025	0.027	0.03
Im6	0.076	0.059	0.137	0.145	0.028	0.058	0.069
Im7	0.416	0.497	0.697	0.339	0.144	0.472	0.572
Im8	0.152	0.18	0.401	0.141	0.027	0.173	0.268
Im9	0.179	0.228	0.38	0.075	0.086	0.228	0.266
Im10	0.06	0.068	0.047	0.065	0.031	0.074	0.085
Im11	0.106	0.085	0.097	0.034	0.028	0.09	0.111
Im12							
Im13	0.109	0.086	0.112	0.051	0.03	0.079	0.075
Im14	0.151	0.206	0.321	0.122	0.042	0.195	0.072

Table 3. Results for MSEE for Imputation Step 2. Results are displayed as MSEE x 100.

	SSR	LIE	LOCF	LME	fPCA	PLS1	PLS2
Im1	0.235	0.223	0.221	0.22	0.22	0.22	0.222
Im2	0.212	0.208	0.22	0.206	0.206	0.207	0.208
Im3	0.189	0.185	0.186	0.184	0.184	0.185	0.185
Im4							
Im5	0.185	0.183	0.185	0.179	0.179	0.182	0.184
Im6	0.216	0.208	0.214	0.202	0.202	0.205	0.208
Im7	0.225	0.221	0.232	0.217	0.217	0.218	0.219
Im8	0.233	0.234	0.249	0.222	0.222	0.226	0.23
Im9	0.223	0.22	0.234	0.213	0.214	0.216	0.218
Im10	0.202	0.2	0.21	0.193	0.192	0.197	0.2
Im11	0.18	0.178	0.18	0.177	0.177	0.178	0.178
Im12							
Im13	0.181	0.179	0.181	0.179	0.179	0.179	0.179
Im14	0.205	0.202	0.208	0.201	0.201	0.202	0.203

Table 4. Results for AIC for Imputation Step 2. Results are displayed as AIC /10^4.