Analysis Approaches for Wearable Device Data

Patrick Hilden

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Pulic Health
in the Mailman School of Public Health

COLUMBIA UNIVERSITY

2021

© 2021

Patrick Hilden

All Rights Reserved

Abstract

Analysis Approaches for Wearable Device Data

Patrick Hilden

Wearable devices, which track a subject's activity (e.g. steps, calories, intensity) over time, have become a popular option for research studies which seek to better understand an individual's physical activity in the day-to-day setting. This thesis looks to address three common problems within the wearable device setting; how to address missing data and incomplete wear time, what to do when large outlying values are present, and how many observation days are required to reasonably estimate various activity metrics of interest. Given the dense nature of observations from such devices, functional data analysis (FDA) provides a natural framework for analysis, and we seek to address the first problem related to missing data by leveraging generalized functional principal components analysis (GFPCA). In addressing the second problem related to outlying values, we leverage both FDA and the novel principal component pursuit (PCP) approach, which has seen limited application within the field, to separate on observed functional value into low-rank, sparse, and error component functions. Finally, using a rich longitudinal data set, we provide insight into the third problem regarding what is an appropriate study length, utilizing the framework of measurement reliability which has been often applied in the activity data setting. Our results suggest that leveraging FDA methods can provide more accurate estimates of activity during periods of nonwear then current approaches, and that in the presence of large outliers more robust estimates of underlying activity and outlier presence can be determined by combining FDA methods and those of PCP. Finally, within our longitudinal cohort we show that current guidelines regarding

the number of days necessary to achieve a reasonable measurement reliability are inaccurate, and often underestimate the true number of days required.

Table of Contents

Dedication iv

Introduction 1

Chapter 1: A Functional Data Approach for Improved Weartime Correction 4

1.1 Introduction 4

1.1.1 Motivation 4

1.1.2 Dataset 6

1.2 Linear Model (LM) Adjustment 7

1.3 Functional Data Analysis (FDA) Adjustment 11

1.3.1 Literature for FDA 11

1.3.2 FDA-Based Weartime Correction 13

1.4 Numeric Studies 15

1.5 Application 17

1.6 Discussion 21

Chapter 2: Robust Decomposition of Accelerometer Data through Functional Principal Component Pursuit 24

2.1 Introduction 24

2.2 Technical Review 28

2.2.1 Functional Principal Component Analysis (FPCA) 28
2.2.2 Principal Component Pursuit 30
2.3 Methods 32
2.3.1 Functional Principal Component Pursuit (FPCP) 32
2.3.2 Tuning Parameters 34
2.4 Simulations 36
2.5 Real World Data Analysis 40
2.6 Discussion 44

Chapter 3: How Many Days are Needed? Measurement Reliability of Wearable Device Data 46
3.1 Introduction 46
3.1.1 Motivation 46
3.1.2 Dataset 48
3.2 Measurement Reliability 51
3.3 Methods 53
3.4 Results 56
3.4.1 Estimate Variability 56
3.4.2 Aggregate Reliability Estimation 57
3.5 Discussion 59
3.6 Conclusions 61

Conclusion 63

References 69

Appendix A: Similarities in Adjustment Approaches 70

A.1 Analytic Results 70

A.1.1 Derivation of Constrained Regression 70

A.1.2 Equivalence in the Absence of Other Covariates 72

A.2 Simulation Results 73

A.2.1 Simulation Design 74

A.2.2 Average Correlation Matrices 75

A.2.3 Activity Coefficient Estimates 76

A.2.4 Weartime Coefficient Estimates 77

Dedication

To my wife Kelly, whose unending love and support carries me through each day.

Introduction

Accelerometers and other wearable devices are now common place in research studies which look to understand habitual patterns of daily activity for subjects under study. This thesis aims to address three key problems which are common to such studies including; what is to be done in the presence of missing data and device non-wear, how to handle large outlying values, and finally the number of days required to accurately capture the habitual activity of a subject.

Chapter 1 focuses on addressing the common issue of device non-compliance or non-wear when utilizing activity monitors within a research study. With subjects typically monitored for several days, failure to wear the device during a portion of the observation period is common. In order to accurately estimate underlying habitual activity, an estimate of likely activity must be incorporated within such periods. A current approach to weartime correction within this setting utilizes a single average activity estimate and imputes this value during all missing data periods and for all subjects. In practice, diurnal activity patterns suggest that the values utilized should incorporate the time at which the missing data was observed, and further the observed activity of a subject when the device was worn. Utilizing the methods of functional data analysis (FDA) and generalized functional principal components analysis (GFPCA), we develop a bootstrap approach to estimating likely activity profiles during periods of device nonwear. Using a real world data set, we show that the given approach which leverages the time period during which missing data was observed, and the observed activity during device wear periods, results in an improved estimate of underlying activity during nonwear periods.

In Chapter 2 we turn to the issue of identifying and accounting for large outlying (sparse)

values observed within physical activity data. Given a goal of identifying the underlying habitual physical activity patterns of a subject, large outlying values which do not fit the common activity patterns across subjects, can greatly impact the estimate of an individual's underlying activity. Principal component pursuit (PCP), which was originally proposed in the computer vision space, can be used to identify outlying objects from the fixed background of an image. Our work looks to adapt the framework of PCP to the functional data setting by combining functional principal components analysis (FPCA) with PCP in an iterative process. Our suggested approach separates each observed functional profile $y_i(t)$ into separate component functions representing low-rank underlying activity profiles, large sparse outliers, and random noise deviation. The resulting low-rank estimates are robust to the sparse outlying values, and using both a simulation study and real world data analysis, we show that these estimates are more accurate and reasonable than those of standard FPCA which does not differentiate between random noise variability and large outlying values. Further, the suggested approach provides a meaningful way to identify time points in which outlying values are present, a beneficial property for many other research questions for which these outliers may be of primary interest.

Chapter 3 focuses on how many observation days are required in a research study in order to accurately estimate habitual activity metrics like average daily sedentary time or moderate/vigorous activity minutes. Within this Chapter, we turn our focus away from analyzing the individual activity profiles $y_i(t)$ and instead focus on the daily activity estimate y_i. Current guidelines suggest that 3-5 days of observation are adequate to reliably estimate average activity metrics among adults. Accordingly, 7 day protocol lengths have become conventional. These guidelines were built using the concept of measurement reliability, and utilizing data from a study with a fixed observation length (e.g. one week), the single measurement reliability of y_i is estimated, and the Spearman-Brown prophecy formula used to generalize the single measure reliability to the number of observation days which should be averaged in a future study to achieve a pre-determined level of reliability. This reliability approach was originally developed within the field of psychometrics, and we show that the underlying assumptions of the prophecy formula are often unmet within the physical ac-

tivity data setting. Using data from a long term followup study, we further provide a sense of the variability in single-measure reliability estimation, and subsequently compare the prophesied reliability to the actual reliability observed given a study protocol length with a varying number of days. Our results highlight that considerable, and yet under reported, variability is present in estimating reliability in physical activity studies, and that current guidelines tend to underestimate the number of observation days required to achieve a given reliability of interest.

Chapter 1: A Functional Data Approach for Improved Weartime Correction

1.1 Introduction

1.1.1 Motivation

The increased use of wearable devices to monitor physical activity in research studies is largely attributable to their low cost, ease of use, relative accuracy, and flexibility, allowing for constant activity monitoring in the day-to-day setting [1]. Frequently, studies using these devices seek to generate summary measures of physical activity (e.g. step counts, minutes of moderate/vigorous activity) over a common time domain across all participants, which are interpretable and can be used either as a predictor or outcome. However, measurement outside of controlled settings typically leads to issues of compliance with device wear, and subsequently to periods of nonwear. Therefore, the ability to properly estimate an individual's physical activity during periods of nonwear is essential to developing robust measurements that can be used to properly evaluate the research questions of interest. Methods which can accurately adjust for nonwear by estimating activity over the common time interval of observation are increasingly in demand.

One popular method to adjust for differing amounts of nonwear uses a linear model [2]. In this method, data are aggregated over time to produce, for each observation day, the total observed activity and the total observed weartime. Then, using the data from all observation days, a linear model assesses the relationship between wear time and activity. For each observation day with incomplete activity data, the relationship between weartime and activity described by this model is used to extrapolate complete activity data from observed activity data. Put differently, this method performs weartime correction by adjusting observed activity with nonwear present to a level mimicking full wear. Several studies have standardized activity in the presence of nonwear using this approach [3, 4]. The corrected activity is often used as a covariate in subsequent models

assessing the effect of activity on an outcome of interest. Alternatively, instead of correcting for different weartimes, it is possible to adjust for both observed activity and weartime when assessing the effect of activity on outcomes [5, 6, 7]; we discuss the close connection between weartime correction and adjustment in Section 1.2 and Appendix A.

The linear model approach to weartime correction does not account for the time of day at which nonwear was observed, effectively assuming that activity is constant over the observed time period. This assumption conflicts with known diurnal patterns in physical activity [8], and ignores the possibility that nonwear may occur more often in times of low activity. The linear model approach can lead to poor estimation of activity as a consequence. A more detailed explanation of the assumptions that underlie the linear model approach to weartime correction and their impact is given in Section 1.2.

Intuitively, an approach which accounts for diurnal patterns of activity and the time period for which nonwear was observed should result in an improved estimate of activity during these periods. Functional data analysis (FDA) [9, 10, 11] provides a natural framework for such an approach. Methods in FDA regard a continuous trajectory of activity over the course of the day as a single unit of observation. These methods incorporate the association between activity and time over the course of the observation period. Approaching the weartime correction problem from this perspective allows for direct incorporation and modeling of the temporal structure inherent in physical activity; leverages information across observation days in determining shared patterns in activity; and uses observed data to estimate activity during nonwear in a way that may better account for the potential relationship between wear likelihood and activity. Several studies have emphasized the applicability of FDA methods to accelerometer data [12, 13, 14, 15]. Section 1.3.1 reviews the FDA literature as it applies to our analysis and a detailed technical description of our method is provided in Section 1.3.2.

We evaluate our method for weartime correction using real data in two ways. First, numerical studies presented in Section 1.4 artificially introduce periods of nonwear into complete data and compare our method to the popular linear model approach; results suggest that our method offers

improved accuracy in estimating daily step counts. Finally, Section 1.5 provides an analysis using real world data from a randomized control trial, and highlights differences between the two approaches.

1.1.2 Dataset

Although the techniques we propose can be used broadly, this work is motivated by a 12-month randomized control trial conducted at Columbia University Irving Medical Center with a primary aim of elucidating the bi-directional relationship between stress and physical activity [16, 17]. A total of 79 participants were included in the study, each of whom was asked to wear an accelerometer device (Fitbit Flex) on their wrist each day for approximately one year. This device recorded minute-by-minute measurements of activity. During the same one-year period, participants were prompted daily via smartphone text-message to report their perceived levels of stress, and whether or not they had exercised for at least 30 minutes that day.

After six months of observation, participants were randomly assigned to either a control or intervention group. All participants were given a report summarizing their typical activity and stress levels during the previous half year. Additionally, those in the intervention group also received information on one to three factors that were associated with their likelihood to exercise, which were tailored to each participant based on his or her activity and responses to daily surveys during the first six months of observation [17].

Our analyses focus on total activity (in the form of step counts) in the 14-hour period from 8:00am to 10:00pm. As a preprocesing step, any two-hour period in which no steps were observed was considered to be nonwear [18]. Days with fewer than ten hours of observed weartime between 8:00am and 10:00pm were considered invalid and are excluded from all subsequent analyses [19], resulting in a minimum and maximum weartime of 10 and 14 hours, respectively. After identifying days that met our inclusion criteria, a total of 15,497 valid days of observation were included. When applying the FDA approach, observed step counts were aggregated to 5 minute epochs, by summing the steps within an epoch, in order to reduce computational burden.

Using data from the Fitbit device, our daily activity outcome of interest is the total number of steps in the 14-hour window defined above. We are particularly concerned with the choice of method for weartime correction, rather than identifying non-wear periods themselves, and our analyses explore the difference between our proposed method and the existing linear model approach.

1.2 Linear Model (LM) Adjustment

The linear model has become a standard for weartime correction when the goal is to produce a single value for activity over the course of a day. The first step in this approach is to determine the total amount of wear time and the total number of steps during that time, for each observation day. Using data from all valid observation days a linear model is then fit to describe the association between wear time and total activity; this produces a line where the slope indicates the expected increase in activity for each additional hour of weartime. For every observation day, incomplete data are assumed to consist of activity at the average rate suggested by the model slope. Put differently, to estimate activity over the full observation period, an observation day's residual from the linear model is added to the expected total activity for an observation day with complete data. The resulting estimated activity is effectively a carrying-forward of observed activity to the maximum observable wear time in a way which is parallel to the observed population activity / wear time trend.

Panel A of Figure 1.1 provides an illustration of the linear model weartime correction using a subset of days selected from the motivating data. In order to emphasize how the LM method adjusts activity estimates during periods of nonwear, we interpret the resulting wear time correction in the context of the original accelerometer data. Panel A shows the slope of the model in this example is 305, implying an imputation of 305 steps per hour of nonwear for every observation day. Panels B and C in Figure 1.1 show the activity profiles from the individual days of observation noted as 1 and 2 within panel A, respectively. The black line in these panels represents the observed activity over the course of the day, aggregated into 5 minute epochs, while gaps in activity identify nonwear.

Panel B shows a day in which there was a large amount of activity over 10 hours of wear time, while panel C shows a low activity day over 10.4 hours. The blue line in each figure represents the epoch-level imputation value (i.e. 305 steps per hour, divided among twelve 5-minute epochs per hour). Under the LM approach, both days use the same imputed value during periods of nonwear despite large differences in observed activity and nonwear patterns.

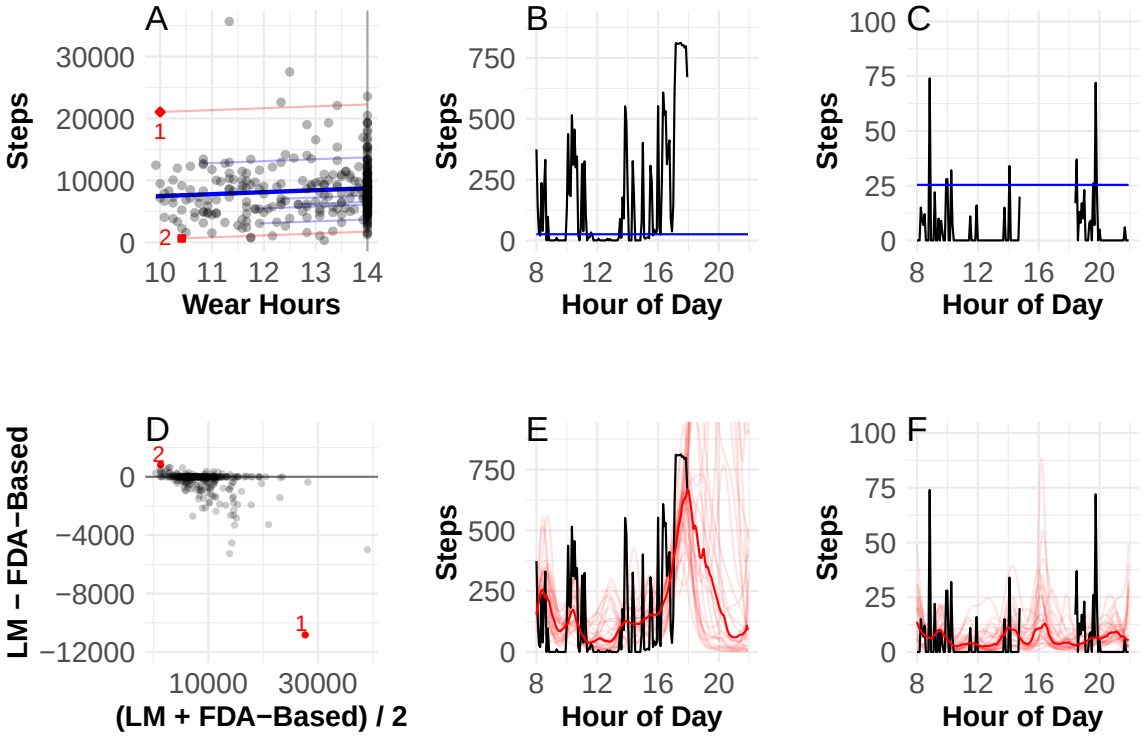


Figure 1.1: Illustration of the linear model and FDA-based approaches to weartime correction in a subset of days from the motivating data. Panel A shows total steps against weartime, with a regression line overlaid in blue, and two observation days highlighted in red. Panel D is a Bland-Altman plot showing the difference in LM and FDA-Based total imputed steps, against the average of the two methods. The center and right columns show activity profiles for the observation days noted as 1 and 2 in the left column, respectively. In panels B and C, the imputation value obtained from the LM method is shown as a blue line. In panels E and F, the activity profile estimated by the FDA-based approach is shown in dark red and estimates from the individual bootstrap samples are shown in light red.

Figure 1.1 makes two main drawbacks in the LM adjustment approach clear. First, the method ignores the time of day in which nonwear was observed, effectively imputing the same activity

value during all nonwear times. However, diurnal variation of daily activity implies that differences in activity exist based on time of day, with individuals tending to be more active during the daytime and less active in the early morning and late evening, an issue not addressed by the LM approach. Second, it uses the same activity rate for each observation day despite obvious differences in activity across observation days. This is particularly concerning for days consisting of either very low or high activity combined with a high rate of nonwear, resulting in likely over- and under-estimation of underlying activity, respectively.

The LM adjustment is used as one step in a data analysis: corrected step counts would typically be treated as a predictor in a subsequent model for some outcome of interest. Another common analysis method bypasses weartime correction and instead includes both observed steps and weartime as covariates in a regression for the given outcome. These methods are, in fact, very closely related. In linear regressions of an outcome on (1) corrected activity or (2) observed activity and weartime, coefficients associated with corrected activity and observed activity will be identical when no other covariates are included in the models. When additional covariates are included, the model that includes corrected activity is a constrained version of a model that regresses on observed activity and weartime. This constraint formalizes the assumption that weartime affects the outcome only through activity when there are no omitted variables which confound the association of weartime with either activity or the outcome. Simulations under a range of scenarios – including settings where that assumption is false – suggest that the two modeling approaches will produce very similar associations between outcome and activity in practice; see Appendix A for additional details.

When constructing linear regression models for an outcome of interest, weartime correction and adjusting for weartime as a covariate are similar. This connection isn't obvious, however, and we find weartime correction and subsequent modeling more conceptually straightforward: a single weartime correction model is implemented, and corrected activity values can be used across a range of outcome models. We therefore note the connection between approaches but focus on the LM weartime adjustment to provide context for the methods we develop in later Sections.

1.3 Functional Data Analysis (FDA) Adjustment

Our aim is to approximate the true activity patterns during periods of nonwear, and use these approximations in combination with the observed data to provide an improved estimate of total daily activity. Intuitively, we use available data from observation days to quantify patterns of activity over time, and then use each day's observed data to determine the extent to which these specific patterns are present for a given observation day. This Section outlines our proposed wear time correction approach, with Section 1.3.1 providing a technical overview of FDA methodology and its application to accelerometer data, and Section 1.3.2 detailing the proposed approach for wear time correction based on the resulting FDA activity estimates.

1.3.1 Literature for FDA

For intensely longitudinal data – that is, data which are observed intensely over a specified interval – traditional tools for longitudinal data analysis may be insufficient for the complexity and richness of observed data. Instead, methods which assume intensely longitudinal data arise from underlying curves $y_i(t)$ and treat these curves as the unit of analysis can more flexibly model the shape and structure of rich repeated measurements. FDA is the branch of modern statistics that provides the conceptual framework for treating intensely longitudinal data, such as those arising from wearable devices, as curves. FDA has a broad literature, and in the following we review the most pertinent techniques.

Functional principal components analysis (FPCA) [9, 20, 21, 14] is a method to estimate the smooth patterns of activity that underlie observed data over the course of a day. To do this, these methods take advantage of information from the entire sample to estimate shared patterns of activity. FPCA methods are akin to traditional methods for principal components analysis (PCA) in that complex, high-dimensional data can be described using a relatively small collection of patterns, effectively reducing the dimension of the data; functional PCA is distinct in that it uses the temporal ordering of data in a way that usual PCA does not. Given shared patterns (functional

principal components), individual daily outcome data are expressed as the combination of these components and day-specific scores which indicate the extent to which individual components are present for a given observation day [22].

Our motivating data are best described by the growing literature on non-Gaussian functional data, with the number of steps per minute assumed to arise from a Poisson distribution for count data. Generalized FPCA (GFPCA) models extend FPCA for continuous outcomes to discrete data using a framework based on generalized linear models [23, 14]. More formally, in GFPCA the underlying mean function of interest for observation day i is given by:

$$E(y_i(t)) = \mu_i(t) \tag{1.1}$$

$$g(\mu_i(t)) = \alpha(t) + \sum_{k=1}^{K} c_{ik}\phi_k(t) \tag{1.2}$$

The observation day specific mean function $\mu_i(t)$ is comprised of a population mean $\alpha(t)$ common to all days, and observation day specific deviations. These deviations are represented by the set of K functional principal components $\phi_k(t)$ arising in the data and observation day specific scores c_{ik} for each of the components, which are assumed to have mean 0, variance $\sigma^2_{c_k}$, and be independent across observation days. The function $g()$ is a known link function between the discrete observation scale and continuous modeling scale; here we use a log link function, as in a usual Poisson generalized linear model.

The components $\phi_k(t)$ capture the main patterns in activity observed in the data, and once determined, can be plotted to get a sense of the variability in activity trajectories they represent, and how they differ over time. The scores c_{ik} capture how much of these patterns are present in a given day, with greater absolute values reflecting a larger effect of a given component in the observed activity trajectory. The components are determined from information across all of the observed activity profiles, and the scores depend on the data present in individual curves.

To this point, we have used continuous activity profile functions as our response of interest to convey the temporal structure in FDA's conceptual framework. However, our data is made up

of a collection of individual observations from these profiles collected over a discrete grid. That is, we observe the epoch-by-epoch measurements of activity from the accelerometer, but we seek to estimate the smooth underlying functions from which these measurements were generated. In practice, the unknown quantities in (1) and (2) are estimated by expanding the overall mean $\alpha(t)$ and component functions $\phi_k(t)$ using a basis expansion over the grid of observed time points.

The resulting model contains several parameters which must be estimated, including the basis function coefficients associated with the previous expansions, as well as the observation day specific scores c_{ik}. In practice, there are several ways to estimate the components and scores in the previous model [24], and for our approach we adapt the methods developed by Backenroth et al. [25].

1.3.2 FDA-Based Weartime Correction

GFPCA has traditionally been used as as means of dimension reduction and applied to data in which little or no missingness is present. In contrast to its historical use, our primary goal is to use GFPCA to generate more accurate reconstructions of activity during periods of nonwear by leveraging information from the observed patterns in activity across all observation days. We then will use these estimates of activity during nonwear in conjunction with observed data to provide an improved estimate of total daily activity.

Estimated activity profiles contain uncertainty in estimated principal components $\phi_k(t)$ and scores c_{ik}; when combined with an exponential link function, even modest over-estimates of these can produce extreme fitted values. We propose to address this numeric instability in GFPCA by producing a collection of estimated activity curves to reflect this uncertainty, and aggregating across this collection using a robust summary. Specifically, we use the bootstrap approach to obtain a collection of estimates that reflect the sampling variability in model parameters and summarize the resulting collection by taking the median at each epoch [26]. This aggregate profile estimate is robust to the effect of any individual profiles within the collection. During periods of nonwear the summary activity profile can be used as an estimate of the underlying activity.

This procedure, which we refer to as “FDA-based weartime correction", is defined as follows. Given n independent observation days of activity data which include complete and incomplete days, we generate a total of b bootstrap samples. These bootstrap samples are constructed by sampling n days from the full data with replacement. The bootstrap samples therefore differ from the original data and from each other, and mimic the variability that underlies the full sample. For a given bootstrap sample we then fit the GFPCA model and determine $\phi_k(t)$, the functional components for that bootstrap sample. Scores c_{ik} are estimated for each observation day in the original cohort, including those which were in the bootstrap sample and those which were not, yielding estimates $\hat{y}_i(t)$ for all observation days. This process is then repeated for each of the b bootstrap samples, and produces a collection of estimated activity profiles for each day in the full cohort that incorporates the uncertainty in the GFPCA process. From this collection, a single estimated profile is determined by calculating the median predicted value in each epoch across the b estimated profiles. Using the FDA-based estimate and the observed data, we determine a weartime corrected estimate of total daily activity by taking the sum of the observed steps and the FDA-based estimated steps during periods of nonwear. In this way, we produce a weartime adjusted estimate of total daily activity for each day which accounts for individual observed patterns in activity and day-specific estimates of activity during nonwear.

Panels E and F in Figure 1.1 illustrate this approach. In both panels, each individual light red line is the estimated profile based on components ϕ_k of a single bootstrap sample, and the dark red line is the FDA-based weartime correction estimate determined by taking the median value across the $b = 30$ bootstrap profiles at each epoch. As expected, the 30 profiles differ substantially from one another, with considerable variability during periods of nonwear, and the noted numerical issues in GFPCA produces some individual estimates that are very large. The overall FDA-based estimates are, however, robust to these infrequent outliers and provide a reasonable extrapolation of activity data within periods of nonwear.

1.4 Numeric Studies

In order to evaluate the accuracy of the LM and FDA-based adjustment methods, we compare the methods in a setting in which the true activity is known. To do so, we artificially introduce nonwear into data where the true activity is known; this nonwear is added under two different scenarios described below. We then apply both the LM and FDA-based methods and compare them using the median absolute prediction error (MAPE) between the true activity and predictions made by both methods.

In the motivating data, there are a total of 6,989 days of observation in which there was complete wear between 8am and 10pm and 8,508 days that had at least some nonwear; we refer to these as complete and incomplete data cohorts, respectively. Using these data as the starting point for this numerical comparison, we imposed nonwear periods on complete data in one of two ways. The first scenario is designed so that the nonwear patterns are similar to those observed elsewhere in the full dataset. The second scenario is designed to include additional informative nonwear, with observation days with less activity in the morning more likely to have “nonwear" in this period. Specifically, in the first scenario we randomly selected a pair of days from the complete and incomplete data cohorts, and imposed the nonwear periods from the latter onto the former. In the second scenario, we started in the same manner but additionally determined the average activity between the hours of 8 and 10am for each day drawn from the complete data cohort. Using weights proportional to the inverse of the average activity in this time, we selected 20% of days drawn from the complete data cohort and set all activity between 8 and 10am to be nonwear. Below, we refer to studies under the first scenario as representative nonwear and those from the second as informative nonwear. Both scenarios produce data that can be analyzed using weartime correction methods, and where the results can be compared to the truth.

We generate datasets with artificial nonwear consisting of 100 total days which were randomly sampled across all subjects, and apply the LM and FDA-based methods. These produce weartime corrections for each day, which are compared to the true activity values using the MAPE. This

same process was repeated 100 times for each scenario.

Figure 1.2 shows the results of this numeric exercise under both scenarios for introducing nonwear. In both scenarios we see considerable improvement in overall activity estimation using the FDA-based method, with a distribution of MAPE values across the 100 simulations which are both smaller and less variable than that of the LM approach. The results show a reduction of 27.3% (from 552 to 401 steps) and 30.4% (from 707 to 492 steps) in the median MAPE under the representative and informative scenarios, respectively, when using the FDA-based approach as compared to the LM method. The MAPE was lower when using the FDA-based method in 83% and 80% of simulations for the representative and informative nonwear scenarios respectively. The greater stability in MAPE values across datasets when using the FDA-based approach suggests that accuracy for this approach is less sensitive to the characteristics of the sample to which it is applied. The MAPE in the informative nonwear scenario is higher than in the representative nonwear, with this increase being larger under the LM approach.

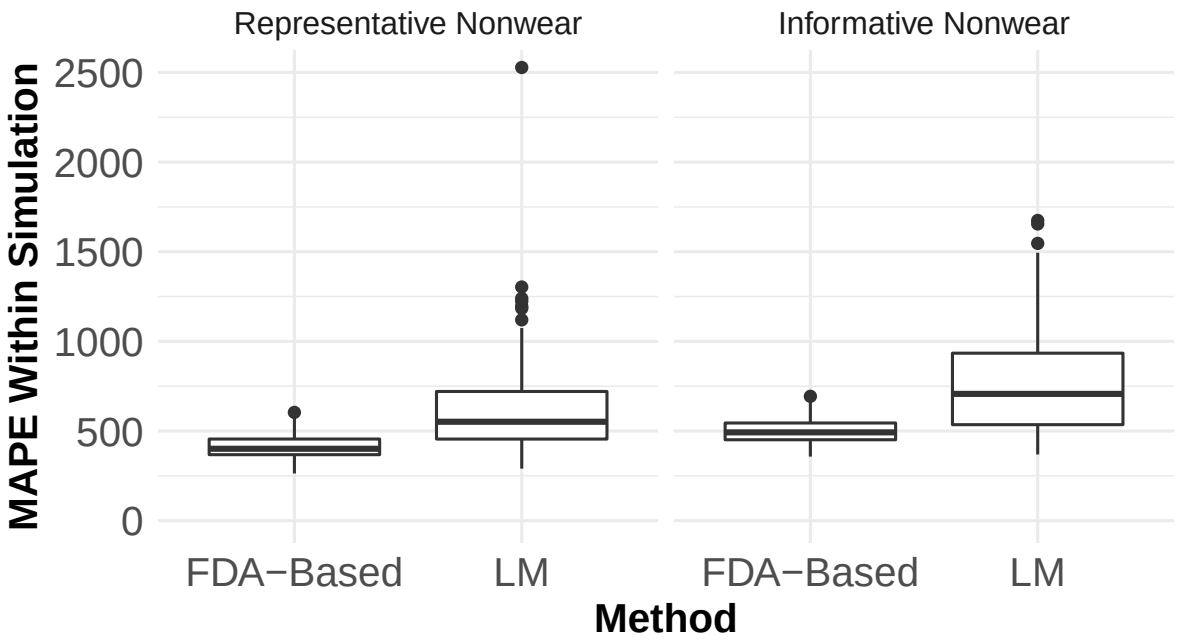


Figure 1.2: Distribution of the median absolute prediction error (MAPE) across replications of our numeric comparison of FDA-based and LM weartime corrections, using the representative and informative nonwear approaches.

1.5 Application

Next we compare the FDA-based and linear model weartime corrections to data where the underlying activity is unknown. Using the motivating data, we selected a single 7-day period during the study in which the majority of participants had observation data. Sixty-six participants are included in the sample, and each participant contributed between 1 and 7 observation days. Thirteen participants without at least one valid day of observation in the given week were excluded from the analysis. This resulted in a total of 327 days of observation, of which 172 days had some missing data between the hours of 8am and 10pm.

Figure 1.3 shows the activity of each individual day of observation across the cohort. The

plot highlights the variability in activity profiles, with some days showing very limited activity throughout the day and others containing sedentary periods interspersed with active bouts of varying length. Gaps in the activity profiles represent periods of nonwear, which tend to occur in the morning or in the evening, with fairly consistent wear during the midday period.

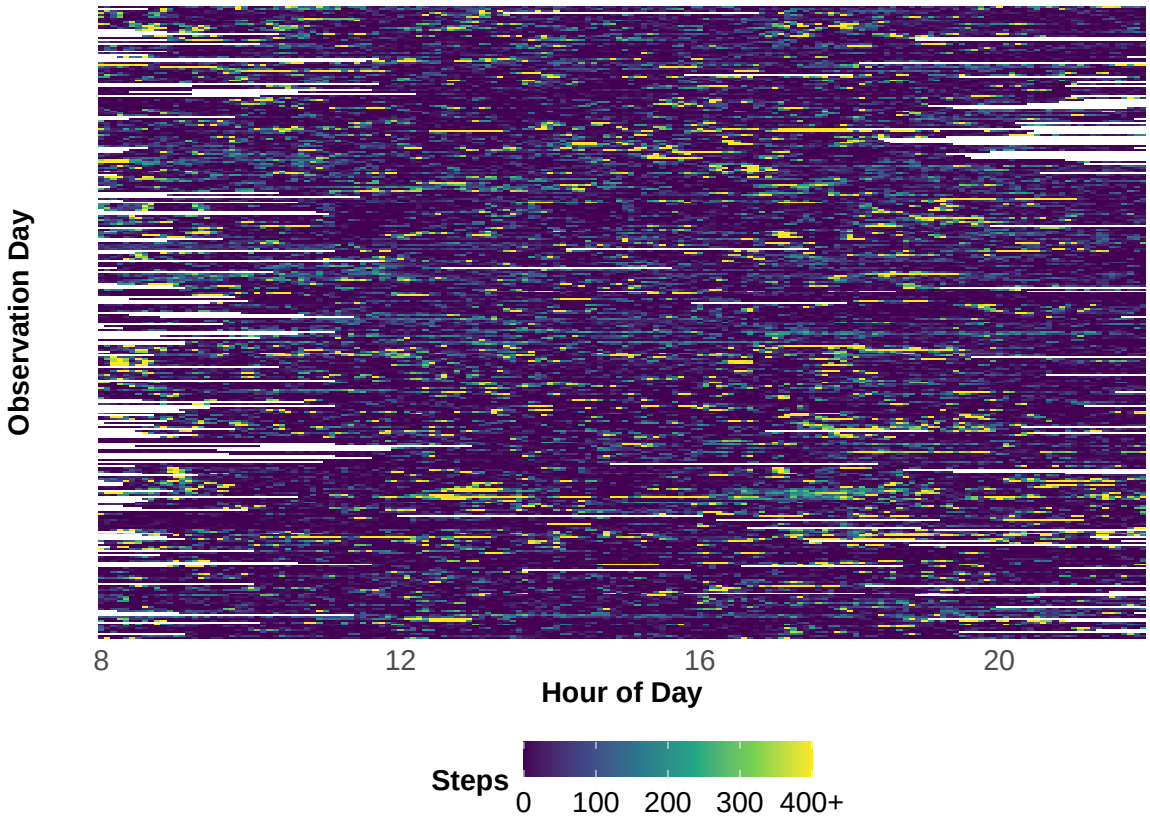


Figure 1.3: Daily activity profiles for 327 observation days among 66 subjects included in our analysis. White regions represent nonwear.

The application of GFPCA to this sample is illustrated in Figure 1.4. The top left panel shows the mean $\alpha(t)$ (black curve) and the mean with the first functional principal component $\phi_1(t)$ added and subtracted (red and blue curves) scaled by the standard deviation of the score values, and exponentiated to lie in the domain of observed data. The top right and bottom left panels repeat this for the second $\phi_2(t)$ and third $\phi_3(t)$ functional principal component. The plot of the

first functional principal component indicates that a major pattern in this sample is an increase or decrease in activity at all times of the day: some subjects are more active and some less. The second functional principal component, meanwhile, identifies a contrast in activity comparing the morning and evening to midday, while the third functional principal component identifies variability in activity throughout the course of the day. Somewhat different components are estimated across bootstrap samples in the application of our method, but the results are qualitatively similar across samples. Three principal components were, in this application, sufficient to estimate activity during non-wear periods without introducing too many unknown parameters.

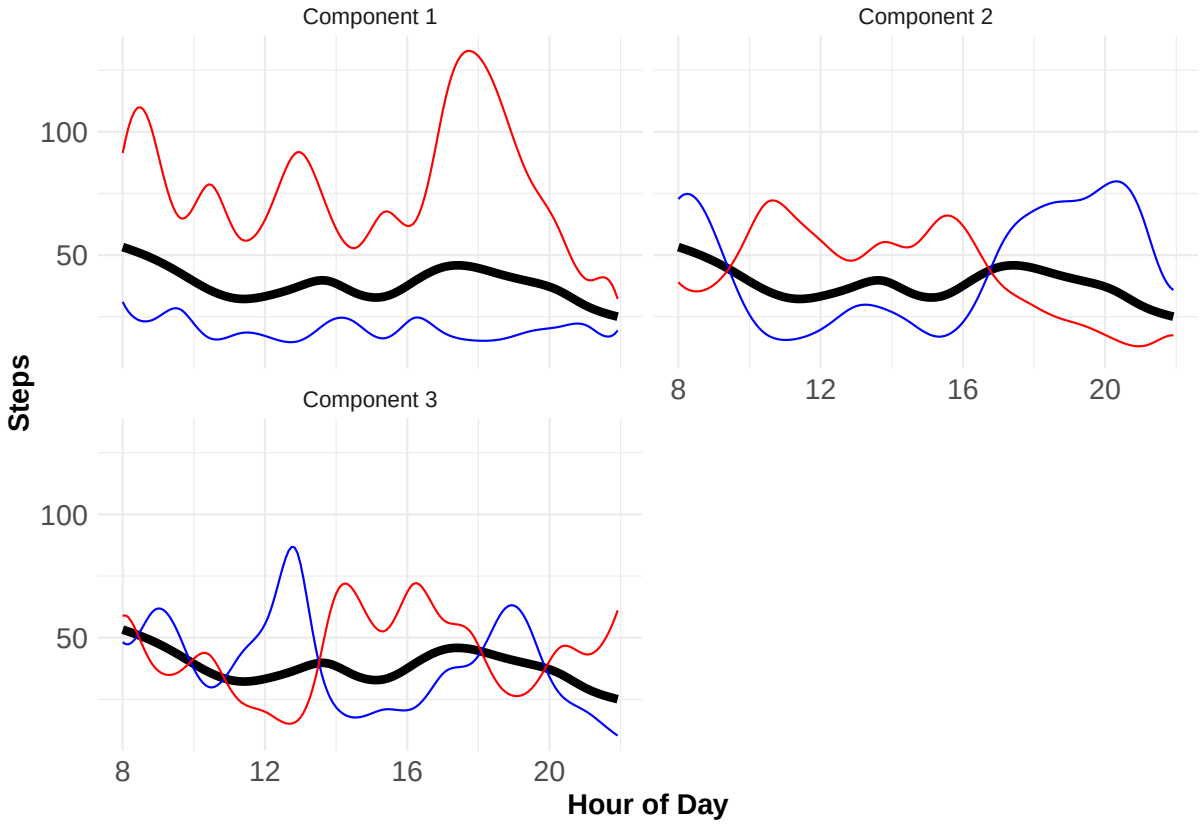


Figure 1.4: Illustrations of the GFPCA approach applied to our analysis sample. Panels show $g^{-1}\left(\alpha(t) \pm \sqrt{\sigma^2_{c_k}}\phi_k(t)\right)$ for all components, with the population mean shown in black.

Panel A of Figure 1.1, as discussed previously, illustrates the application of the linear model weartime correction to these data. In the scatterplot of total steps against wear hours and in the overlaid linear regression line, there is a weak relationship between these and a resulting LM imputation value of 305 steps per hour (approximately 25 steps per 5-minute epoch) of missing data. Panel D of Figure 1.1 shows a Bland-Altman plot comparing the difference in the two approaches against their average, highlighting the same subjects as panel A. Observation days with no non-wear fall on the black line. For days with relatively few observed steps, the LM approach tends to impute more steps than the FDA-based approach; the reverse is true for days with higher observed steps. This trend emerges because the FDA-based weartime correction is informed by the data available for each day, while the LM method uses the same imputation value regardless of available data.

The center and right columns of Figure 1.1 show the activity data from the two days of observation noted as 1 and 2 in the left column. Both observation days contain a high proportion of missingness, but have very different amounts of activity when the device was worn. In panels B and C, the blue line represents the LM imputation value. This is consistently lower and higher than the observed activity in the center and right panels, respectively, illustrating issues that can arise when using the same value for imputation for all observation days, regardless of observed activity.

Panels E and F show the bootstrap fits (light red lines) and the median fitted value (dark red line) arising from the FDA-based approach for the same two subjects. The observation day in panel E contains many periods of high activity, and this is reflected in the estimated activity profile for the non-wear period. Indeed, the estimated evening peak in activity is consistent with observed data and with the functional principal components shown in Figure 1.4. The observation day in panel F, meanwhile, contains persistently low activity, which is reflected in the activity profile estimated by the FDA-based approach.

Comparing panels B and E of Figure 1.1 demonstrates how the LM and FDA-based weartime corrections can produce vastly different estimates for the same observation day. Observation 1 is an outlier – as shown in panel D, for this observation day the methods produce dramatically differ-

ent estimates, with the FDA-based approach estimating approximately 33,000 total daily steps as compared to 22,000 under the LM approach. Unlike the analyses in Section 1.4, we are unable to say which is closer to the truth. However, we suggest that the estimated activity profile in the FDA-based approach is a reasonable, data-driven extrapolation of observed steps. We also note that in another observation day (not shown), this subject accrued roughly 27,000 steps in 12.5 hours of wear, suggesting high step counts are not implausible.

1.6 Discussion

The numeric studies presented in section 1.4 show that the LM weartime correction approach can often lead to considerable inaccuracy in estimation of the underlying activity, and that a FDA-based approach which accounts for the timing of nonwear can result in improved estimation of underlying activity profiles. Further, the results under the informative nonwear scenario show that the FDA-based method is more robust in scenarios in which the likelihood of nonwear is associated with the underlying activity, a situation which is reasonable to expect in many settings.

In addition to providing an overall estimate of total daily activity, the FDA-based approach also provides the benefit of estimating activity profiles over time, which accounts for diurnal patterns of activity, and allow for a more granular estimate of activity as compared to that of the LM-approach which disregards the timing of missing data during the day. The ability to estimate activity profiles throughout the observation period also allows the benefit of evaluating the profile fit during periods of observed activity, allowing for assessment of the predicted accuracy and fit.

The application highlighted in section 1.5 provides additional insight into the performance of both the LM and FDA-based approaches, highlighting the increased accuracy of the FDA-based approach in a setting typical to that encountered by many investigators. The relationship between wear time and activity in panel A of Figure 1.1 provides an example of the poor results which can result from the LM approach. In this example, we see relatively little relationship between wear time and activity, resulting in an imputation of 305 steps per hour of missingness, a value much lower than one would expect. The individual days highlighted in Figure 1.1 show that this

imputation value may differ considerably across the majority of time where activity was observed, and the FDA-based fitted values provide more accurate estimation of activity during periods of observed wear, and more reasonable predictions of activity during nonwear.

While the FDA-based approach provides increased accuracy in the estimation of total daily activity, its application has a number of limitations. Considerable computation time can be required when estimating the functional profiles for each of the individual bootstrap samples, particularly when the number of participants and the grid of observation times is large. Implementation of the FDA-based approach on a computation cluster which can allow for the bootstrap estimation of individual profiles simultaneously, as was used in our analysis, allows for computation of the numeric study results in a matter of minutes, as the overall time to compute all bootstrap samples is reduced to that of computing a single result. Additionally, there remains a risk of poor estimation of the underlying activity profiles during large periods of nonwear, particularly when these occur at the boundaries of the observation time. Examination of the resulting predicted activity profiles is necessary to ensure a reasonable fit during periods of both wear and nonwear. The FDA-approach also requires setting reasonable parameter values in both the GFPCA model fitting, as well as the bootstrap sampling. In our experience we have found that using 20 basis functions and 3 principal components provided reasonable fit in a variety of scenarios for the GFPCA model fitting, while using 30 bootstrap samples was adequate for increasing prediction accuracy and reducing the influence of outlying predictions in any one sample.

Our work focuses on weartime correction at the observation day level. This is often a first step in data processing, after which multiple days are aggregated into a single measure of activity. The impact of this aggregation is difficult to predict. On one hand, the impact of large discrepancies between weartime correction methods may be attenuated in an aggregate measure, especially if those discrepancies are limited to a small number of individual days. On the other hand, Figure 1D suggests that differences between methods may be systematic, and a persistent bias in corrected steps would appear in aggregate measures. In either case, the construction of an aggregate measure of activity depends on weartime corrections in individual days, and the proposed approach provides

a more accurate correction than widely-used techniques.

Future work will seek to develop an approach with increased numeric stability, reducing the impact of outlying predictions and the need for a bootstrap approach. Additionally, incorporating the correlation in activity between observed days on the same participant may lead to improved results when applying an FDA-based approach.

Chapter 2: Robust Decomposition of Accelerometer Data through Functional Principal Component Pursuit

2.1 Introduction

Accelerometers are widespread in research settings where participants' physical activity in their day-to-day lives is of interest [1]. The popularity of these devices stems from the richness of the data they provide, often at low cost and minimal burden on study participants. Daily activity profiles that often consist of epoch-level measurements of physical activity can be obtained through continuous monitoring of each participant in the observation period; epoch lengths of one minute are common, although finer resolutions are possible. These profiles provide insight into the timing and intensity of physical activity, and are made up of recognizable elements. A few population-level patterns are typically sufficient to understand major components of activity: the level of overall activity in the daytime hours, for example, or the contrast between activity in the morning and in the afternoon. In addition to the broad trends these patterns describe, it is common to observe intermittent spikes of high or low activity in an activity profile that are distinct from more typical minute-to-minute variation, and may arise from bouts of exercise, brief engagement in strenuous activity, or an irregular period of inactivity. Methods that can estimate common activity patterns and also accurately identify short bursts of very high or low activity are therefore necessary to provide a complete understanding of physical activity.

This understanding for the composition of physical activity profiles is illustrated in Figure 2.1. The top row shows activity count trajectories over the course of one week for a single subject, with activity counts binned within 5 minute epochs. These panels suggest clear diurnal patterns of activity, with relatively little activity during the nighttime and early morning and higher activity in the daytime hours. There is day-to-day variability in the general intensity and timing of activity

and epoch-to-epoch fluctuation around these patterns. Importantly, there are also infrequent but obviously outlying activity count values observed on some days and times. Although typical values of activity counts rarely exceed 10,000 there are brief spurts of activity counts which are between 10,000 and 30,000. The second and third rows of Figure 2.1 make the separation between the broad patterns and large but infrequent deviations from these patterns explicit. Finally, the last row in Figure 2.1 shows remaining minute-to-minute variability, so that within each column the full profile in the top row is the sum of the elements in the remaining rows.

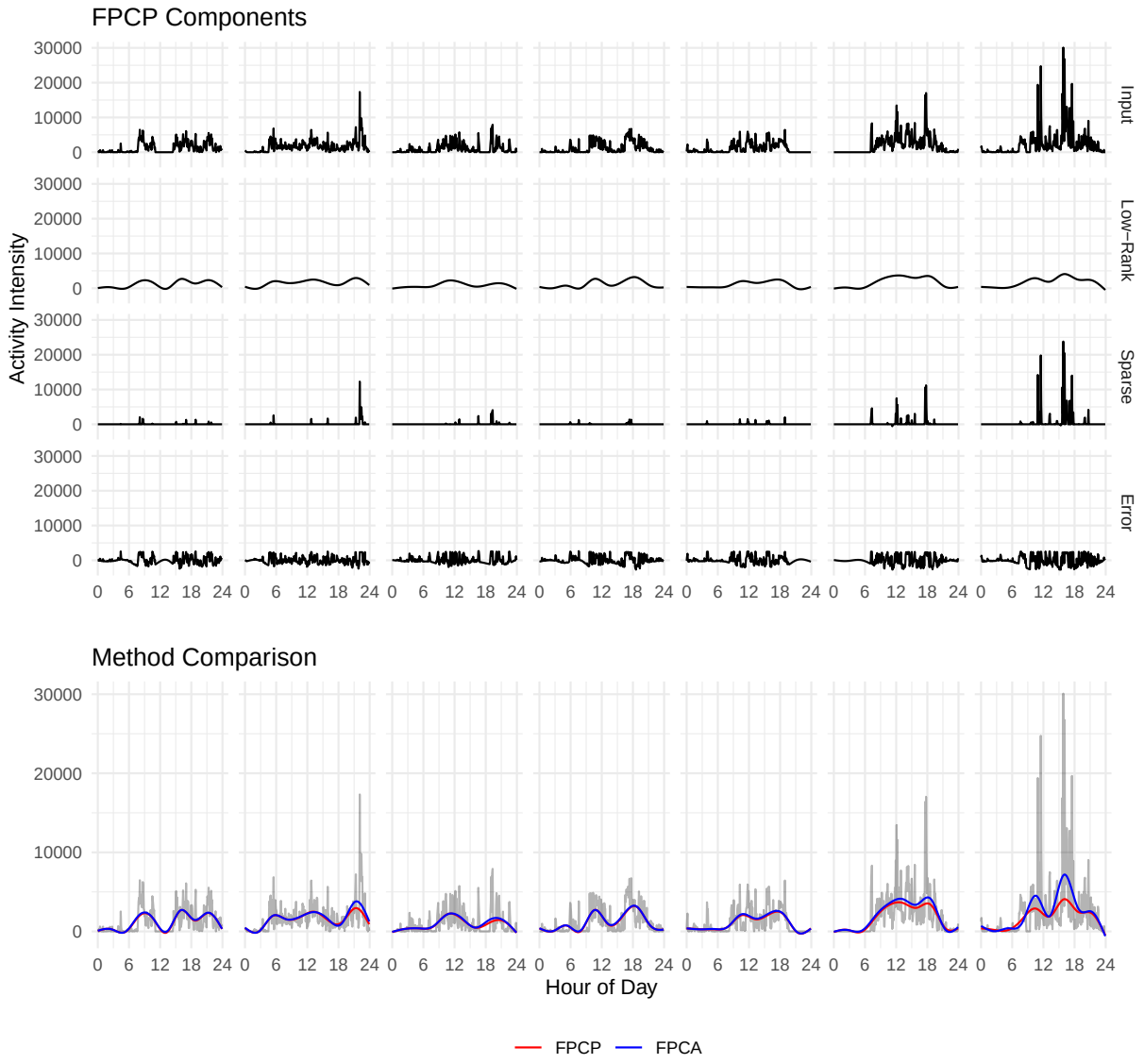


Figure 2.1: Real data FPCA and FPCP method comparison. The top row presents a single week of activity counts for a subject within the motivating data. The 2nd, 3rd, and 4th rows provide the FPCP breakdown of the observed activity into low-rank, sparse, and error components, respectively. The bottom row compares the FPCA and low-rank FPCP fits for the observed activity.

Our novel approach to partitioning activity profiles, which generates the separation seen in Figure 2.1, draws on two bodies of work. First, because activity profiles are observed over a

specific time domain and represent an underlying continuous process they are naturally viewed from the perspective of functional data analysis (FDA)[9, 10, 11]. Indeed, the development and use of FDA methods in accelerometer research has become more common in recent years [12, 13, 14, 15, 27, 28]. In particular, functional principal components analysis (FPCA) has been used to identify the main directions of variability in accelerometer data, and for each activity profile the extent to which these directions are present. Although it has been widely used for activity data, FPCA does not distinguish between moderate minute-to-minute variation and large deviations from underlying patterns. That is, FPCA fails to identify bouts of very high or very low activity and can therefore be sensitive to these outlying values.

We also build on principal component pursuit (PCP), which was originally proposed within the computer vision literature and is sometimes referred to as robust principal components analysis. PCP is an effective approach for segmenting the changing elements from a collection of images with a fixed background; a common application is the detection of moving objects in a digital video where the frame is unchanging [29]. In order to identify outlying aspects in a set of images, PCP separates images, represented as data matrices, into low-rank and sparse components. The low-rank component expresses directions of variation in the fixed portion, or background, of an image. The sparse component, meanwhile contains the infrequent outlying or irregular deviations from the background. PCP achieves the separation of low-rank and sparse components through an iterative algorithm, and recent extensions have included an error matrix that is distinct from the low-rank and sparse components to capture low-variability noise. PCP has seen applications in a number of other areas, including video surveillance, topic and keyword identification in text processing, and fault detection within engineering processes [30, 31, 32, 33, 34]. Despite the broad appeal, PCP has not been widely used in the statistical literature or adapted to a functional data setting.

This paper introduces functional principal component pursuit (FPCP), a novel combination of FPCA and PCP, that is intended to separate observed activity profiles into an element explained by low-rank patterns describing main directions of variation, an element containing infrequent out-

liers, and the remaining noise. This method has several advantages over existing approaches for dimension reduction in the context of wearable device data. FPCA does not distinguish between small variations of random noise and large outlying values, and the resulting FPCA component estimates and scores can vary considerably in the presence or absence of outliers. PCP isn't directly applicable to functional data where temporal information is relevant, and won't produce smooth estimated activity profiles. Finally, the identification of sparse outliers themselves may be of interest, for example as a way to identify exercise bouts, and an approach which can separate the elements of observed activity profiles may be useful in these settings. The bottom row of Figure 2.1 shows a comparison between the fitted trajectories of the FPCA and our proposed PCP based approach, highlighting the sensitivity of both approaches to outliers.

Section 2.2 provides a technical overview of FPCA and PCP, and Section 2.3 outlines our proposed approach. Section 2.4 provides a simulation study which highlights the improvements in the proposed approach, as compared to traditional FPCA, while Section 2.5 conducts an analysis of the motivating data, highlighting the differences in results when applying the proposed approach and FPCA. Finally, we provide a discussion of our method and some future directions for research in Section 2.6.

2.2 Technical Review

Our goal is to develop FPCP, an alternative to traditional FPCA that is robust to outlying values by adapting PCP to a functional data setting. In Section 2.2.1, we briefly review FPCA to establish notation, and in Section 2.2.2 we review existing PCP methods for matrix factorization.

2.2.1 Functional Principal Component Analysis (FPCA)

FPCA is among the most well-established analytic techniques for functional data. In contrast to PCA methods for non-functional data, FPCA approaches use the structure inherent in functional data to improve estimation, often through explicit smoothing steps. FPCA methods allow one to identify the principal directions (components) of variation within the data, and further use

these components as a means for dimension reduction. There are a number of available methods for FPCA with distinct assumptions, implementations, and strengths [35, 20, 36, 37]. For concreteness, we present an overview of the specific method used in later Sections, but other FPCA methods could be used as well with no fundamental changes to the proposed FPCP methodology.

Let $x_i(t)$, $1 \leq i \leq n$ and $t \in [0, 1]$, be a functional observation with mean $\mu(t) = E(x_i(t))$ and covariance surface $\Sigma^X(t, t') = \text{Cov}[X_i(t), X_i(t')]$. Also let the eigen decomposition of $\Sigma^X(t, t')$ be given by $\sum_{k=1}^{\infty} \lambda_k \phi_k(t) \phi_k(t')$, where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots$ are the eigenvalues and $\phi_k(t)$ the orthornormal eigenfunctions of the decomposition. Generally, only a few eigenfunctions and eigenvalues are needed to capture the main directions of variation; this property makes FPCA appealing as a tool for dimension reduction. Based on the preceding decomposition, the truncated Karhunen-Loeve expansion for curve $x_i(t)$ is

$$x_i(t) = \mu(t) + \sum_{k=1}^{K} c_{ik} \phi_k(t) \tag{2.1}$$

where $c_{ik} = \int_t x_i(t) \phi_k(t)$ are observation-specific scores with mean 0 and variance λ_k representing the extent to which a given eigenfunction $\phi_k(t)$ contributes to an observed curve.

Conceptually, functional data can be observed without error over a compact interval; in practice, however, data are observed with noise over a finite grid. Let $y_i(t) = x_i(t) + \epsilon_i(t)$ and assume $\epsilon_i(t)$ is distributed normally with mean zero and variance σ^2. From observations of curves $y_i(t)$ of discrete time points, FPCA seeks to estimate the mean, components, and scores in equation (2.1). To carry out this estimation, we first estimate a smooth mean using penalized splines fit to observations pooled across subjects. Next, to obtain an estimate $\hat{\Sigma}^Y(t, t')$ of the covariance surface, we smooth a method-of-moments estimate of the covariance using bivariate penalized smoothing splines after removing the main diagonal elements. We obtain the eigen decomposition of the smoothed covariance surface in order to estimate λ_k and $\phi_k(t)$. Lastly, we estimate scores c_{ik} by treating them as random effects in a mixed model with known eigenfunctions [20, 21].

2.2.2 Principal Component Pursuit

Principal Component Pursuit (PCP) has become a popular approach for separating the low-rank and sparse portions of an observed data matrix, but does not account for the serial correlation structure present in functional data. Therefore, in contrast to Section 2.2.1 we now focus exclusively on data observed with no functional structure.

Given a data matrix Y, and assuming observations were observed with error, PCP seeks to separate the matrix into low-rank (L), sparse (S), and error (E) structures:

$$Y = L + S + E.$$

This decomposition is accomplished by minimizing an objective function that combines squared error loss on the reconstruction with penalties on both L and S:

$$\min_{L,S} \quad \|Y - (L + S)\|_F^2 + \tau_1 \|L\|_* + \tau_2 \|S\|_1. \tag{2.2}$$

Here, $\|Y - (L + S)\|_F^2$ the squared Frobenius norm of the difference between the observed data and the combined low-rank and sparse matrices (i.e. error), $\|L\|_*$ is the nuclear norm of L, and $\|S\|_1$ is the l_1-norm of S. The nuclear norm $\|L\|_*$ is the sum of the singular values of L (or roots of the eigenvalues of LL^T); while not a direct measure of rank, this provides a measure of the complexity in L and allows for a convex solution where a direct penalty on rank does not. Including the l_1 penalty $\|S\|_1$ in Equation 2.2 allows for some large values in S but encourages the majority of this matrix to contain zeros [38]. The L matrix, and its associated penalty, seek to balance the rank of the approximation to the observed data with the degree of reconstruction error; the matrix S, meanwhile, contains large errors that would be influential in the Frobenius normal but is zero otherwise.

The tuning parameters τ_1 and τ_2 control the amount of attenuation in the singular values of L and the degree of sparsity in S by balancing the goodness of fit for Y using L and S and controlling

for their relative complexity. For larger values of τ_1, singular values are more penalized which induces a reduction in the variability of the low-rank estimate, while the opposite is true for smaller values of τ_1. Similarly, the sparsity is controlled by τ_2: as τ_2 increases, fewer elements are included in the sparse matrix S, and vice-versa for smaller values of τ_2. While theoretical results for PCP suggest plausible values for these tuning parameters, they can be chosen using cross validation; see Section 2.3.2 for details relevant to our methods.

For fixed tuning parameter values, the minimization in Equation 2.2 can be performed using an iterative approach which alternates between the estimation of the low-rank and sparse structures. In each cycle through the iterative algorithm, one first obtains the singular value decomposition $U\Sigma V$ of $Y - S$ (the current low-rank estimate L). The resulting singular values Σ are attenuated via a shrinkage and thresholding function defined by

$$S_\alpha[x] = \text{sgn}(x)\max\left(|x| - \alpha, 0\right)$$

where α controls the the degree of shrinkage applied to singular values, and is applied elementwise to Σ. The L matrix is then given by singular value recomposition using the attenuated singular values $S_{\tau_1}[\Sigma]$. Next, given the current estimate of L, S is determined by $S_{\tau_2}[Y - L]$. This process is then iterated until convergence. The complete PCP algorithm is presented in Algorithm 1.

Algorithm 1 PCP Algorithm

1: **initialize** $S_0 = 0, \tau_1 > 0, \tau_2 > 0$
2: **while** not converged **do**
3: compute singular value decomposition ($U\Sigma V$) of $Y - S_k$
4: compute $L_{k+1} = US_{\tau_1}[\Sigma]V$
5: compute $S_{k+1} = S_{\tau_2}[Y - L_{k+1}]$
6: **end while**

2.3 Methods

2.3.1 Functional Principal Component Pursuit (FPCP)

In contrast to traditional dimension reduction approaches for functional data, we pose a conceptual model that contains a specific term for large, outlying deviations from a subject-specific smooth trajectory that can be distinguished from a separate noise term. In particular, we assume

$$y_i(t) = \mu(t) + l_i(t) + s_i(t) + \epsilon_i(t) \tag{2.3}$$
$$l_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$$

where $\mu(t)$ is the a population mean; $l_i(t)$ is a subject-specific deviation from the mean that can be explained by a small number of shared components K via a basis expansion; $s_i(t)$ is an error term that contains sparse, outlying deviations from the smooth subject-specific profile; and $\epsilon_i(t)$ represents remaining residual variation. Functional data are, in practice, observed over finite grids $\boldsymbol{t}_i = \{t_{i1}, t_{i2}, \ldots, t_{iJ_i}\}$ for each subject $i = 1, 2, \ldots, I$. We will use $y_i(\boldsymbol{t}_i)$ to denote the length J_i vector obtained by evaluating the function y_i at the timepoints in $\boldsymbol{t}_i$; similar notation holds for other functions evaluated over vectors.

Our goal is to estimate the low-rank element of equation (2.3) while simultaneously partitioning the remaining variation into sparse, large outlying values and remaining residuals. As in PCP, for given tuning parameters τ_1 and τ_2, we develop an iterative estimation approach consisting of two main steps: first, we obtain a low-rank decomposition of curves after subtracting the sparse error using FPCA; second, we separate deviations from the low-rank fit into sparse errors and remaining noise. These steps are iterated until convergence, which is monitored using a criterion described below.

In the first step, given a current estimate of sparse error functions $s_i(\boldsymbol{t}_i)$, we conduct FPCA on the curves defined by

$$\begin{aligned}\tilde{y}_i(\boldsymbol{t}_i) &= y_i(\boldsymbol{t}_i) - s_i(\boldsymbol{t}_i) \\ &= \mu(\boldsymbol{t}_i) + l_i(\boldsymbol{t}_i) + e_i(\boldsymbol{t}_i).\end{aligned}$$

After removing the sparse error term $s_i(\boldsymbol{t}_i)$, the curves $\tilde{y}_i(\boldsymbol{t}_i)$ will follow the data generating mechanism presumed by the FPCA model structure. We can, therefore, focus on obtaining a low-dimensional representation for the structure in the $l_i(\boldsymbol{t}_i)$ using basis functions shared across subjects. From this, we obtain estimates of $\mu(\boldsymbol{t}_i)$, c_{ik}, λ_k, and $\phi_k(\boldsymbol{t}_i)$. Our implementation uses the FPCA approach outlined in Section 2.2.1, although as noted there other techniques for FPCA which can be used without fundamentally altering our methods.

We choose a large truncation lag K and, as in PCP, attenuate the eigenvalues $\hat{\lambda}_k$ using the shrinkage and thresholding operator $S_{\tau_1}[\hat{\lambda}_k]$. Let $\hat{c}'_{ik} = \hat{c}_{ik}/\hat{\lambda}_k$. The estimated low-rank profiles that result from FPCA and the attenuation of eigenvalues are given by

$$\hat{l}_i(\boldsymbol{t}_i) = \sum_{i=1}^{K} \hat{c}'_{ik} S_{\tau_1}[\hat{\lambda}_k] \hat{\phi}_k(\boldsymbol{t}_i).$$

This explicit shrinkage of eigenvalues reduces the impact of the truncation lag K, instead the tuning parameter τ_1 informs data-driven estimates of eigenvalues and can shrink unnecessary components to zero. The choice of tuning parameter is discussed in Section 2.3.2.

In the second step, given a current estimate of the low-rank functions $\hat{l}_i(\boldsymbol{t}_i)$, we estimate the sparse error functions $\hat{s}_i(\boldsymbol{t}_i)$ by applying the shrinkage and thresholding operator to the difference between the observed function and current low-rank estimate:

$$\hat{s}_i(t_i) = S_{\tau_2}[y_i(\boldsymbol{t}_i) - \hat{\mu}(\boldsymbol{t}_i) - \hat{l}_i(\boldsymbol{t}_i)].$$

The operator S_{τ_2} returns zero where the absolute value of differences between the observed data $y_i(\boldsymbol{t}_i)$ and current low-rank estimate $\hat{l}_i(\boldsymbol{t}_i)$, are less than τ_2. Elsewhere, this returns values that

have been reduced by τ_2, effectively partitioning the observed difference into a unexpectedly large component and a residual. Thus, τ_2 establishes an upper limit for variation that falls within an expected range, and differences between $y_i(\boldsymbol{t}_i)$ and $\hat{l}_i(\boldsymbol{t}_i)$ that exceed this appear, in part, in $\hat{s}_i(\boldsymbol{t}_i)$ and are considered sparse.

The full FPCP approach is provided in Algorithm 2. Given tuning parameters τ_1 and τ_2 and initializing $s_i(\boldsymbol{t}_i) = 0$ for all i, the FPCP algorithm iterates between the two steps described above. This iterative process is repeated until convergence.

Algorithm 2 FPCP Algorithm

initialize $\hat{s}_i(\boldsymbol{t}_i) = 0$, $\tau_1 > 0$, $\tau_2 > 0$
while not converged **do**
    compute FPCA decomposition of $y_i(\boldsymbol{t}_i) - \hat{s}_i(\boldsymbol{t}_i)$
    extract $\hat{\mu}(\boldsymbol{t}_i)$, $\hat{\lambda}_k$, $\hat{\phi}_k(\boldsymbol{t}_i)$, and $\hat{c}_{ik}$ from the FPCA decomposition
    compute $\hat{c}'_{ik} = \hat{c}_{ik}/\hat{\lambda}_k$
    compute $\hat{l}_i(\boldsymbol{t}_i) = \sum_{i=1}^{K} \hat{c}'_{ik} S_{\tau_1}[\hat{\lambda}_k]\hat{\phi}_k(\boldsymbol{t}_i)$
    compute $\hat{s}_i(\boldsymbol{t}_i) = S_{\tau_2}[y_i(\boldsymbol{t}_i) - \hat{\mu}(\boldsymbol{t}_i) - \hat{l}_i(\boldsymbol{t}_i)]$
end while

In the applications below, we define our objective function for convergence as the largest absolute difference in any element in the sparse functions $\hat{s}_i(\boldsymbol{t}_i)$ between iterations and declare that a local minimum has been reached and the method has converged when this elementwise change is $< 10^{-5}$.

2.3.2 Tuning Parameters

The tuning parameters τ_1 and τ_2 respectively control the extent of attenuation of eigenvalues in the low-rank component, and establish a threshold on the magnitude of deviations from the low-rank structure for inclusion as a sparse, outlying value. Previous work for PCP has suggested default values for τ_1 and τ_2 [39].

PCP estimates low-rank structures using a singular value decomposition, and the default value for τ_1 is intended to apply to this framework. For functional data settings where curves are observed over a regular grid on [0,1] that is shared across subjects, this default value can be readily adapted. Specifically, we define τ_1 by

$$\tau_1 = \frac{4 \cdot \sum_i \sum_{\boldsymbol{t}} |y_i(\boldsymbol{t}) - \hat{\mu}(\boldsymbol{t})|}{I \times J} \tag{2.4}$$

where J is the length of the shared grid, and $\hat{\mu}$ is obtained through a direct (unsmoothed) estimate. In this expression, the J term in the denominator reflects a Reimann sum approximation for the integral of eigenfunctions. Our simulations and real data analyses suggest that this is a plausible default value for τ_1 in the setting we consider, although other values may be needed for sparse observations or irregular grids.

The default τ_2 value suggested in earlier works failed to produce reasonable results in our simulations and real data analyses. As such, we determine τ_2 based on cross validation. We randomly select 20% of observed time points from each functional observation to comprise the testing data, and use the remaining observations as training data. We set τ_1 as above and define a grid of candidate τ_2 values. For each τ_2 value we then fit the FPCP method to the training data; note that an FPCA method that allows for missing values is needed for this step. We compute the squared error between the values in the test set and the low-rank fit from the FPCP approach. The results are then aggregated across all of the functional observations to determine the median mean squared error across all values in the test set. The largest τ_2 value which minimizes the previous is thus selected and utilized within the full analysis of the data. In determining the grid of candidate τ_2 values, it is helpful to recall that τ_2 provides a threshold on the magnitude of sparse outliers, in that differences between an observed value and the low-rank estimate larger than τ_2 will appear, in part, in the sparse error term. In this way, a priori intuition about the reasonable size of a sparse value can be incorporated in deciding which τ_2 values are reasonable candidates within the cross validation.

Again, the default τ_1 values assume that all of the functional observations were observed over the same dense, regularly spaced grid of observations $\boldsymbol{t}$. When this is not the case (and when other appropriate defaults cannot be found), a bivariate grid search using cross validation is a possible approach to choosing tuning parameters.

2.4 Simulations

We use simulations to evaluate the ability of FPCP to accurately estimate a low-rank structure in the presence of large, outlying values, and to identify the true outlying values. As part of this, we compare to a direct application of FPCA.

We generate functional observations according to

$$y_i(t) = c_{i1} \cdot cos(t) + c_{i2} \cdot sin(t) + s_i(t) + \epsilon_i(t) \tag{2.5}$$

with scores $c_{i1} \sim N(0, \sigma^2 = 4)$ and $c_{i2} \sim N(0, \sigma^2 = 1)$ representing the extent to which the two primary components comprising the low-rank portion of each curve are represented. Each curves is evaluated on a shared, equally spaced grid $\boldsymbol{t}_i \in [0, 2\pi]$ with 288 individual observation points, in order to represent data taken over a 24-hour period, and summarized into 5-minute epochs. Residual values $\epsilon_i(\boldsymbol{t}_i) \sim N(0, \sigma^2 = 0.25)$ represent random noise added to the observed data. We generate sparse outlying observations in $s_i(\boldsymbol{t}_i)$ by randomly creating between 0 and 4 bouts of outliers in each curve. The number of timepoints in each bout is distributed $unif(5, 10)$, and the value of $s_i(\boldsymbol{t}_i)$ across all timepoints in a given bout is distributed $unif(1, 4)$. For each sample size in $\{50, 100, 200\}$ we generate 100 datasets and fit both FPCP and FPCA to each dataset.

FPCA analyses, including those contained in the iterative FPCP algorithm, are implemented using fpca.sc from the refund R package [40]. In FPCP steps, we set the truncation lag $K = 10$; in our comparison to FPCA alone, we use the default setting that retains enough components to explain 95% of variation. For FPCP, we use the default value of τ_1 provided in equation (2.4), and choose τ_2 using cross validation in a single simulated data set for each sample size, using a grid of 40 equally spaced values between 0.1 and 4. FPCP converged in roughly 6 iterations for datasets containing 50 curves, and roughly 10 iterations in datasets containing 100 and 200 curves.

We quantify the effectiveness of the proposed FPCP method and standard FPCA approach in recovering both the low-rank components $\phi_1(t) = cos(t)$ and $\phi_2(t) = sin(t)$, as well as the individual-level reconstructions $l_i(t)$. In each simulated dataset, estimation accuracy is measured

through the integrated mean square error (IMSE). For the low-rank components we compute

$$\int_t \left(\hat{\phi}_k(t) - \phi_k(t)\right)^2 .$$

For individual reconstructions, we compute

$$\frac{1}{I}\sum_i \int_t \left(\hat{l}_i(t) - [c_{i1} \cdot \phi_1(t) + c_{i2} \cdot \phi_2(t)]\right)^2$$

for each profile i and average this value across all I profiles to obtain an average IMSE. We evaluate the accuracy of FPCP in identifying sparse outliers using the Sorenson-Dice coefficient [41]:

$$\frac{2\sum_i \sum_{\boldsymbol{t}} I\left(s_i(\boldsymbol{t}_i) \neq 0\right) \cdot I\left(\hat{s}_i(\boldsymbol{t}_i) \neq 0\right)}{\sum_i \sum_{\boldsymbol{t}} I\left(s_i(\boldsymbol{t}_i) \neq 0\right) + \sum_i \sum_{\boldsymbol{t}} I\left(\hat{s}_i(\boldsymbol{t}_i) \neq 0\right)}$$

comparing the individual timepoints where outliers were induced, with whether or not any portion of the timepoint was identified within $s_i(\boldsymbol{t}_i)$ of the FPCP approach.

The top row of Figure 2.2 shows generated profiles for a subset of 5 simulated observations from a dataset containing 200 profile curves. For the majority of the observation domain, we see a curve with a moderate amount of noise, while for the subset of timepoints in which outlying values were induced we see larger deviations from the underlying functional observations. Other rows in the top portion of Figure 2 illustrate the decomposition of observed data in to low-rank, sparse, and residual components that results from the application of FPCP. These panels show that FPCP is, in general, accurate in estimating each element. The bottom portion of Figure 2.2 compares the low-rank estimates obtained by both FPCP and FPCA, and illustrates that FPCP is less affected by outlying values and more accurate for estimating the true low-rank structure.

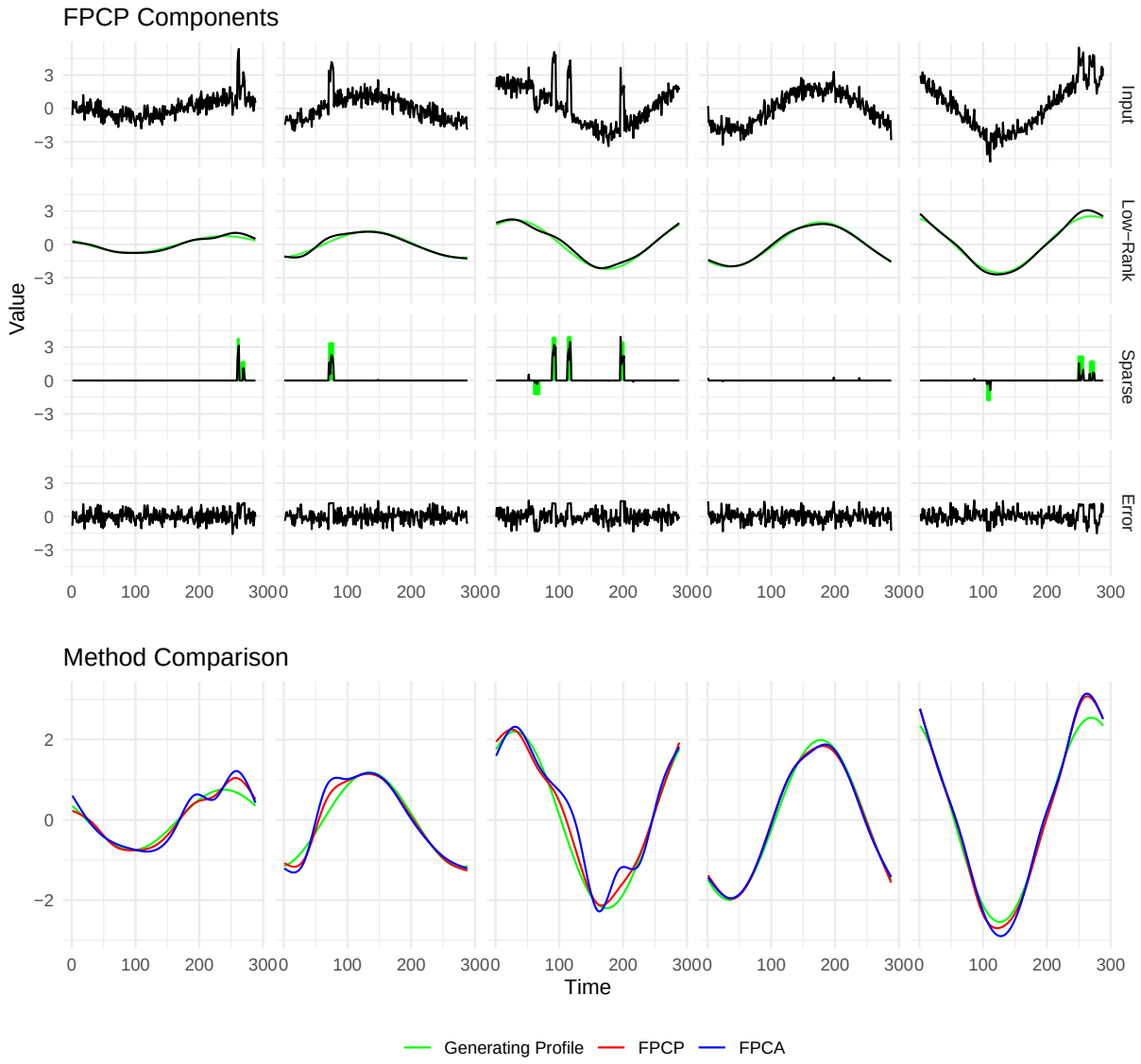


Figure 2.2: FPCA and FPCP method comparison from simulated data containing 200 profiles. The top row presents a selection of 5 simulated profiles. The 2nd, 3rd, and 4th rows provide the FPCP breakdown of the observed activity into low-rank, sparse, and error components, respectively. The bottom row shows the underlying data generating profile, without sparse or error components, and compares the FPCA and low-rank FPCP fits for the simulated profiles.

The top left and right panels of Figure 2.3 show the distribution of IMSE for ϕ_1 and ϕ_2 for

both FPCA and FPCP approaches across the varying sample sizes. For each sample size and both components, we see modest reductions in IMSE for FPCP compared to FPCA, and both methods become more accurate as the sample size grows. The bottom left panel of Figure 2.3 shows the distribution of the mean IMSE for low-rank components for both approaches and all sample sizes. As expected, the FPCP approach provides a substantial reduction in mean IMSE in each sample size, reflecting the robustness to outlying values provided by the method. We see relatively little reduction in the FPCA mean IMSE as the number of simulated profiles increases, while the mean IMSE for the FPCP approach declines at a greater rate as the number of profiles included increases. Within the 200 simulated profile simulation, there was a decrease of 58.2% in the average mean IMSE for the FPCP approach as compared to FPCA. The bottom right panel of Figure 2.3 provides the distribution of the Sorenson-Dice coefficient across the sample sizes. As the sample size increases, we see improved Sorenson-Dice coefficients and reduced variability in the coefficient; this may stem from improved estimation of ϕ_1 and ϕ_2. The mean Sorenson-Dice coefficient across the 100 simulations using 50, 100, or 200 simulated profiles was 0.79, 0.80, and 0.84, respectively, indicating strong agreement between locations in which outliers were induced, and sparse portions identified within $s_i(\boldsymbol{t}_i)$.

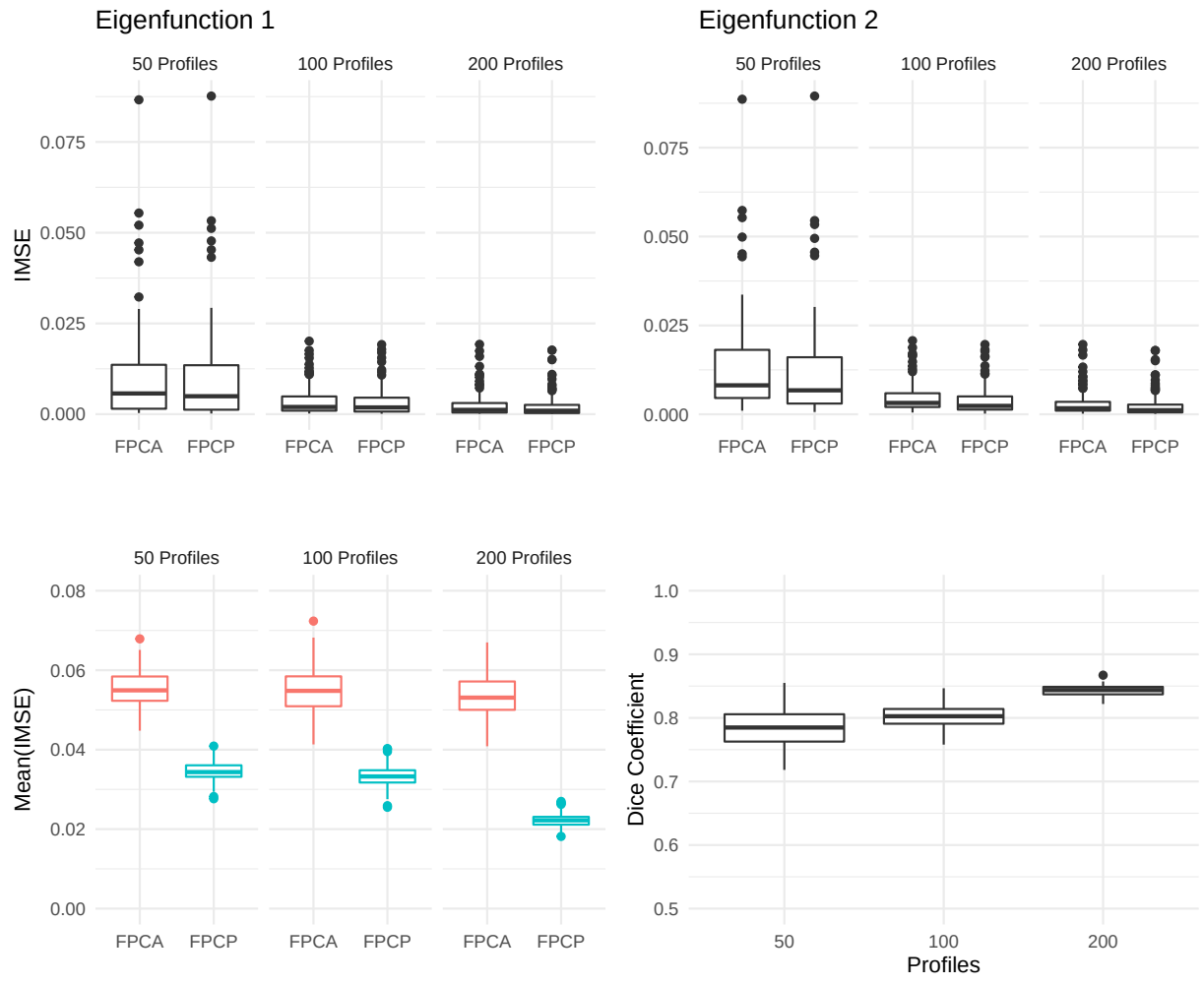


Figure 2.3: Simulation results comparing FPCA and FPCP for sample sizes of 50, 100, and 200 simulated profiles. The top left and right panels compare the distribution of integrated mean square error values across simulations for each eigenfunction. The bottom left panel shows the distribution of the mean IMSE across all profiles within a simulation for each of the various sample sizes, while the bottom right panel shows the distribution of the dice coefficient.

2.5 Real World Data Analysis

We now turn to data from a physical activity study which sought to characterize the relationship between physical activity measurements and adverse events in a cohort of 59 subjects with conges-

tive heart failure [28]. Subjects were instructed to wear an accelerometer device which captured minute-by-minute activity over the course of the day throughout the study, and were monitored between 4 and 9 months. The accelerometer measured activity counts, which is a measurement of the observed activity relative to rest for a given subject [42], were aggregated to the 5-minute epoch level. For the purposes of our analyses, we focus on a single subject who had 329 total days of observation. We excluded any days in which 50% or more of the observation times had activity counts equal to zero, which resulted in an analysis cohort of 282 days of valid observation data.

We analyzed these data using both FPCA and FPCP to contrast the results obtained by these approaches. For FPCP, we set τ_1 to the default value given in Equation 2.4 and we chose τ_2 using cross validation from a grid of equally spaced values from 1,000 to 5,000 by increments of 100. For FPCP, we set $K = 10$ and for FPCA we select K so that 95% of variation is explained, resulting in 10 retained components.

Figure 2.4 provides a lasagna plot of the subject's activity over the 282 days of observation, with a single week highlighted. Data for each day in that week is shown in panels in the top row of Figure 2.1 using lineplots. From these views of the activity data we see clear diurnal patterns of activity, with relatively little activity at night (11pm - 6am), and significantly greater activity during the day time. Additionally, there are several relatively brief bouts of very large activity values, often occurring between 6-9am or 6-9pm. Figure 2.1 in particular shows the existence of sparse outlying activity counts which can be more than 5 times greater than typical activity count values.

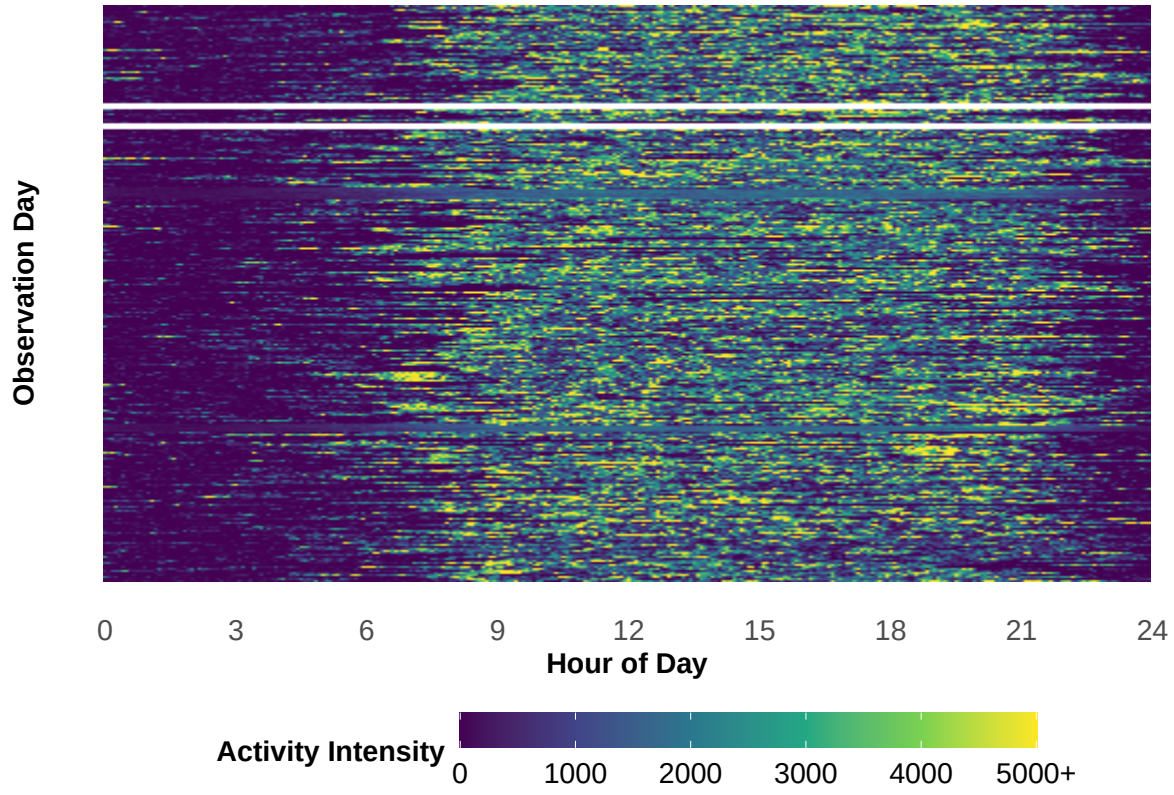


Figure 2.4: Lasagna plot showing the activity counts for each of the 282 observation days (rows) in the motivating data for the selected subjects. The separated rows represent the week of data shown in Figure 2.1

.

The top portion of Figure 2.1 provides a breakdown of the observed daily activity data into low-rank, sparse, and error component estimates obtained through the FPCP approach. The low-rank profiles are smooth estimates of activity counts over time, and are constructed from patterns estimated across all observation days. The sparse components are zero for the majority of observation times, with non-zero values corresponding to periods of notable outlying values in the observed data. Finally, the error component contains remaining residual variation of low magnitude relative to the non-zero values in the sparse term.

The red and blue lines in the bottom panel of Figure 2.1 represent the FPCA and FPCP fits when applying both methods to the full set of observation days. In periods in which the overall activity counts are low (< 5,000) both the FPCA and FPCP methods produce nearly identical estimates of the underlying smooth activity profiles. During bouts of high activity, the FPCP and FPCA methods differ, often notably: FPCP is robust to outlying values, and the low-rank estimate is less affected by bouts of high activity than the estimates obtained by FPCA.

Figure 2.5 displays the first 4 eigenfunctions resulting from the low-rank estimation of the FPCP method. The first eigenfunction represents an overall shift in activity among those who tend to be more or less active throughout the day. The second component represents a shift in activity in the morning (6-9am) or late evening (6-9pm). The third eigenfunction represents a shift in the very early morning starting at 4am, and variability in early and mid afternoon activity, with the 4th eigenfunction representing additional variability throughout the course of the day. As in simulations, these eigenfunctions are broadly similar to those obtained by FPCA. The sensitivity to outlying values in FPCA arises through the estimates of scores on leading components and the lack of attenuation of later eigenvalues.

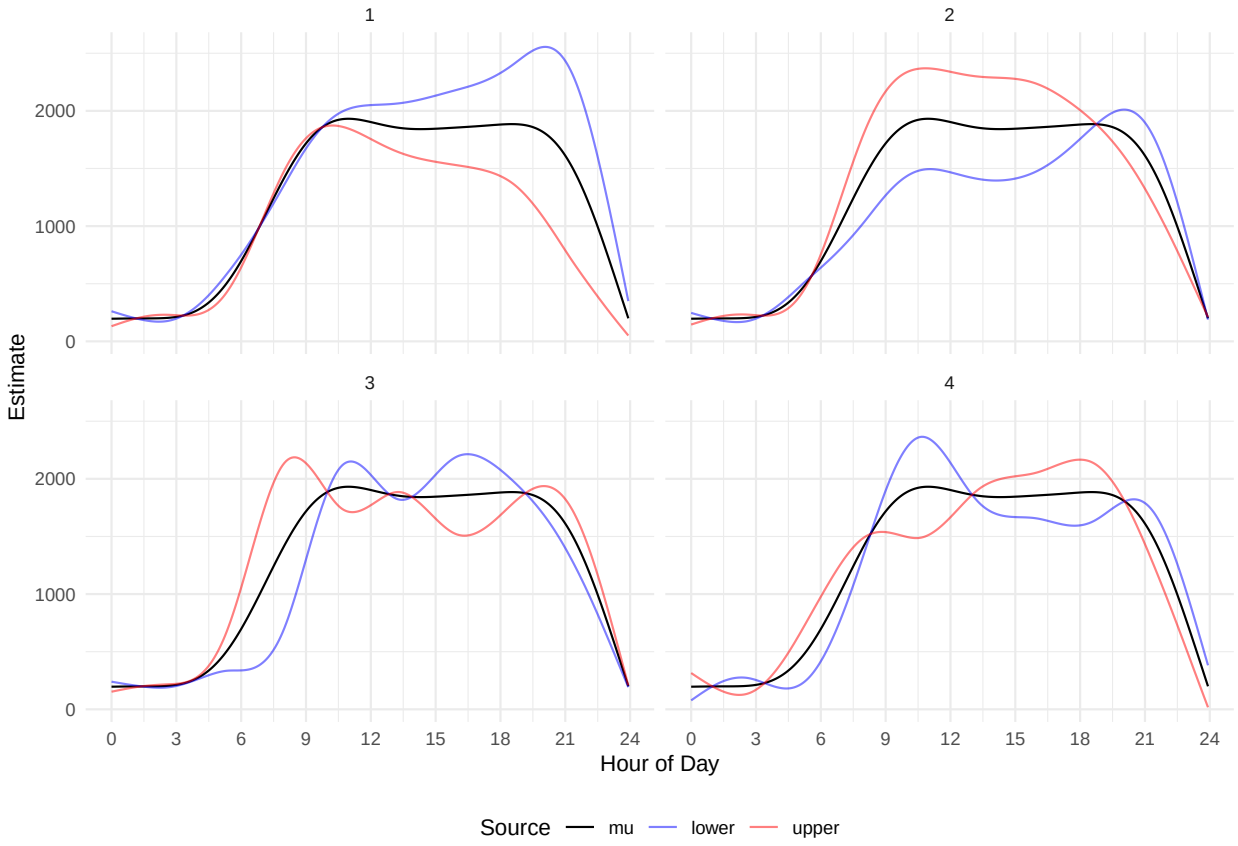


Figure 2.5: Eigenfunctions from the motivating data analysis. Panels show $\mu(t) \pm \sqrt{\lambda_k}\phi_k(t)$ for the first 4 eigenfunctions from the final FPCP fit, with the population mean shown in black.

2.6 Discussion

Dimension reduction is a common tool in the analysis of functional data, both as an exploratory technique and as first step in analysis pipelines. In this manuscript, we have introduced an approach, FPCP, that extends traditional FPCA approaches to reduce sensitivity to outlying values. This has at least two advantages over the direct use of FPCA. First, the obtained low-rank structures are robust to outliers and more reflective of the true underlying profile. Second, we add potentially relevant information to the decomposition by explicitly identifying outliers. Both components may be useful in both exploratory and formal techniques, and this partitioning may be reflective of the

true data generating mechanism in many contexts. In our motivating setting, for example, a low-rank structure that reflects habitual activity and identified bouts of high activity may enter into models for health outcomes separately.

Our work combines elements of functional data analysis and image processing in a unique and innovative way. Simulations and an analysis of wearable device data suggest that the proposed approach is successful in estimating a low-rank structure and identifying outlying values. Convergence of the proposed iterative algorithm is generally fast, requiring only a few iterations. The most time consuming element in each iteration is FPCA; adopting fast alternatives to the approach implemented here would further reduce the computational burden. While our method is developed with physical activity data in mind, the approach is general and our implementation is publicly available.

There are several directions for refinements of FPCP and future work. We have developed methods that are general, and allow sparse or irregular observations across subjects. That said, our motivating context involves functional data observed over dense and regular grids. We suspect that adaptation to other settings will be direct – the biggest challenge we foresee is tuning parameter selection – but we have not explored performance for sparse data here. Our approach does not share information across the functional domain in identifying outlying values, and doing so could improve estimation of these terms. Finally, we note that alternative approaches to ensuring robustness are possible, although techniques like median or quantile regression remain underdeveloped for functional data.

Chapter 3: How Many Days are Needed? Measurement Reliability of Wearable Device Data

3.1 Introduction

3.1.1 Motivation

Studies that use wearable devices often produce summary metrics like sedentary, light (LPA), and moderate to vigorous activity (MVPA) time, in order to measure habitual daily physical activity for study participants. Recognizing that there is some day-to-day variability in activity within subject, individual daily observations are usually aggregated by averaging across days to obtain more accurate subject-level measurements of each physical activity metric. While it is clear that a greater number of days of observation will lead to more stable subject-level activity measurements, it is not well known how many days of observation are “enough" – that is, what number of days will be sufficient to produce an aggregate measure that reflects actual habitual activity levels. Previous studies have framed this as a question of measurement reliability, and used related methods to estimate the number of observation days needed to meet a reliability threshold [43, 44, 45, 46, 47]. However, it is not obvious that reliability is appropriate for the physical activity setting, and no studies to our knowledge have used long term follow-up data to determine the actual reliability provided by data collected in suggested observation periods.

Measurement reliability was originally introduced in the field of psychometrics, but has since seen applications in a variety of other areas [48, 49, 50, 51, 52]. The most common framework assumes that each subject has an underlying true value, with measurements differing from this true value due to independent, identically distributed random deviations. This assumes, for example, that each participant has his or her own true habitual sedentary time, and each day’s measurement

is a completely random deviation from that true time. Given a set of subjects with multiple measurements on an outcome of interest (e.g. total sedentary minutes for each subject and day over a week), reliability is defined as the ratio of subject-to-subject variability to total variability, ranging from 0 to 1 [53]. Reliability is high when the magnitude of random deviation is small relative to the differences between subjects, and vice-versa, with a generally accepted standard for good reliability being 0.8.

Somewhat counter-intuitively, reliability describes a feature of a single measurement – what information a single day's sedentary time would contain if no other days were available – even though it requires multiple observations to estimate. In the context of physical activity, it is further necessary to consider the reliability of an observation obtained through aggregating single measurements. Given an estimated reliability for a single observation based on observed data, the Spearman-Brown prophecy formula is intended to determine the number of observations per subject which, if averaged, would result in a predetermined level of aggregate reliability. Current recommended practice for the determination of habitual physical activity via accelerometry, based on applying the prophecy formula in a variety of previous studies, suggest aggregating measurements from 3-5 days among adults, and 4-9 days among children [19, 54]. Accordingly, 7-day accelerometer protocols have become conventional in the field [55, 56, 57, 58, 59].

There are a number of limitations to providing an estimated number of days needed to achieve an aggregate reliability of 0.8 based on the prophecy formula. First, the number of days derived from the prophecy formula is an estimated value, but variability in this estimate has been under reported. Second, the underlying statistical assumptions surrounding the estimation of reliability, including the independence of measurement days within subjects, and the shared variance of deviations across subjects may often be unmet in practice. Finally, just as multiple observations are needed to understand the reliability of a single observation, it is necessary to obtain multiple independent aggregate measures to establish their reliability; without that follow-up, it is unclear if multi-day averages achieve their prophesied reliability.

In order to better understand the variability associated with estimating reliability in practice,

the corresponding results of applying the prophecy formula, and the actual reliability of aggregate measures, we present an empirical study of single-measure and aggregate reliability utilizing data from a long-term followup study that collected device-measured physical activity data for 365 consecutive days. We find that there is a substantial variability in the estimated number of days needed to achieve a target reliability. More importantly, our results suggest that the process using the prophecy formula underestimates the number of days needed to obtain a reliability of 0.8, suggesting that current best practice recommendations regarding accelerometer protocol lengths may need further evaluation.

Section 3.1.2 describes the data from a long-term followup cohort which motivates our investigation. Section 3.2 provides a detailed overview of measurement reliability and the prophecy formula, and Section 3.3 overviews the proposed methods for evaluating the use of reliability within the physical activity setting. Section 3.4 highlights the results of our investigation, and we conclude with a discussion of the results and their implications in Section 3.5.

3.1.2 Dataset

The motivating data for our evaluation comes from a 12-month randomized control trial which was conducted through the Columbia University Irving Medical Center which sought to better understand the bi-directional relationship between physical activity and stress. Seventy-nine subjects were included in the study and were asked to wear an activity monitor (Fitbit Flex) on their wrist each day for approximately one year [16, 17].

Our analysis focuses on the typical waking hours of 8:00am to 10:00pm. Any two-hour period where no steps were observed was considered to be nonwear [18], and we required a minimum of 10 hours of wear time over the 14 hour observation period for a given day to be valid and considered for inclusion in our analysis. For each subject, we excluded the first 14 days of observation as a run-in period to allow for the possibility that an individual's activity may differ during this time due to device reactivity [60]. After discarding the run-in period, we selected the next 60 valid days of observation for each subject; this provides a period that is long enough to obtain multiple

non-overlapping observation windows to study reliability, but short enough that habitual activity levels are not affected by substantial behavioral changes. Subjects were enrolled in the study over a six month period, minimizing potential confounding due to seasonal differences. Ten subjects who did not have at least 60 valid days of observation during the study were excluded. For each valid day we used the device step counts per 1-minute epoch (e.g. step cadence) to determine the total number of sedentary (steps = 0 for a given 1-minute epoch), light ($0 < \text{steps} < 100$ for a given 1-minute epoch), and moderate/vigorous (steps ≥ 100 for a given 1-minute epoch) minutes of activity. We used step counts in our definition of intensity, rather than device-produced intensities, due to lack of information regarding the device intensity algorithm, the implausibility of some of the produced values, and its corresponding validity for distinguishing physical activity intensity (particularly for sedentary and LPA time). Our choice of thresholds uses an established approach to defining intensity based on per minute steps counts to infer step cadence. [61, 62]. Adjustment for non-wear was done via the linear model weartime correction [2].

The resulting analysis data set consisted of 69 subjects each having 60 days of observation. The median age within the cohort was 29 years (range = [20, 58]) with 40 (58%) being female. The mean sedentary and LPA hours per day were 9.7 (SD = 1.4) and 3.9 (SD = 1.4), respectively. The median MVPA minutes per day was 17 (IQR = [6, 32]). The median number of days on study required to achieve 60 valid days of observation was 81 (IQR = [72, 98]). The resulting dataset is illustrated in the top row of panels in Figure 3.1, which shows each individual's activity metrics over the 60 days of observation. This highlights that there is both day-to-day variation within a subject and subject-to-subject variation, both of which contribute to measurement reliability. The majority of subjects had between 8 and 12 hours of sedentary time a day and, as expected, no systematic changes within subjects were observed.

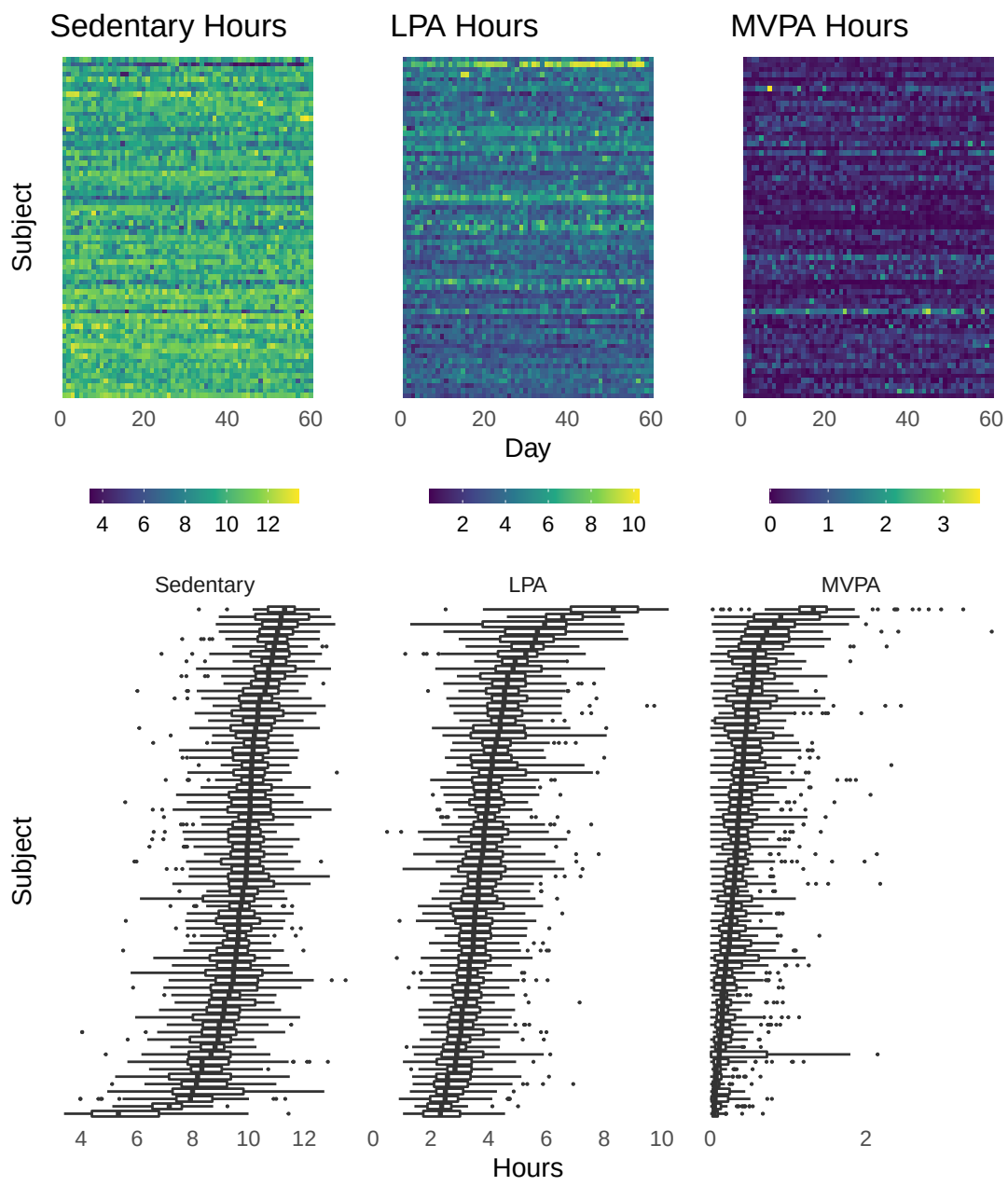


Figure 3.1: Activity metric distributions in the motivating data set. The top row of panels shows the heatmaps for each metric and subject over the 60 days of observation. The bottom row shows the distribution of sedentary, light, and MVPA time for each subject over the course of the study, sorted based on the median value for each metric separately.

3.2 Measurement Reliability

We begin with a discussion of the statistical framework for measurement reliability as it has been used in physical activity. Measurement reliability is a richly researched topic, and our goal is not to provide a complete literature review. Instead, this discussion is intended to make key concepts concrete, and will clarify which statistical assumptions of a particular reliability definition may not be valid in the current setting.

Reliability quantifies the degree of similarity of observations within a subject, with the understanding that measurements are taken in unchanging conditions. It is assumed that individual measurements are the combination of a true subject-level value and random deviations from that. The model assumed to generate an observed measurement X is

$$X = T + e \tag{3.1}$$

where T is the underlying true value intended to be measured and e is a random deviation from that value. It is assumed that T has mean μ and variance σ_T^2 while e is a mean zero residual with variance σ_e^2; the residual has mean zero under the assumption that the measurements are unbiased for the true value. Further, it is assumed that T and e are independent. The reliability of a single measurement R_1, then, is defined as the percentage of overall variability which can be attributed to true subject-to-subject differences:

$$R_1 = \frac{var(T)}{var(X)} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}.$$

This framework emphasizes that, conceptually, reliability depends on partitioning individual measurements into true scores and noise.

In practice, estimating reliability depends on data with multiple measurements per subject. We assume data arise from

$$X_{ij} = T_i + \epsilon_{ij} \tag{3.2}$$

with subjects $i = 1, \ldots, n$ and replicates $j = 1, \ldots, J$. In addition to the assumptions that accompany Equation 3.1, we assume that errors are uncorrelated within subjects across replicates j. Given such data, measurement reliability R_1 can be estimated using a mixed effects model with random intercepts for each subject. The model produces estimates $\hat{\sigma}_T^2$ and $\hat{\sigma}_e^2$, which in turn can be used to estimate the reliability $\hat{R}_1$.

It is well known that the average of a set of independent measurements of the same quantity provides a more accurate estimate of the underlying true value of interest. Given independent replicate measurements for each subject, the aggregate reliability of the average $\bar{X}_i = \frac{1}{J} \sum_{j=1}^{J} X_{ij}$ is given by

$$R_{\bar{J}} = \frac{var(T_i)}{var(\bar{X}_i)} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2/J}.$$

As the number of replicates J increases, there is a corresponding increase in aggregate reliability $R_{\bar{J}}$. This relationship can be framed in terms of the reliability of a single measurement via

$$R_{\bar{J}} = \frac{JR_1}{1 + (J-1)R_1}.$$

The preceding is known as the Spearman-Brown prophecy formula, which relates the single-measurement reliability R_1 and the J-measurement average reliability $R_{\bar{J}}$. The prophecy formula has been utilized in physical activity studies to obtain an estimate $\hat{J}$ of the number of days J that are necessary to produce a given level of aggregate reliability, most typically $R_{\bar{J}} = 0.8$, for activity metrics like sedentary, LPA, and MVPA time.

The definition of reliability and subsequent application of the prophecy formula makes assumptions that may not hold in practice. First, it is assumed the observations X_{ij} made within and across subjects are independent. Second, it is assumed that the residual variance σ_e^2 is constant across measurement days and the same for all subjects. However, it is unlikely that activity across consecutive days are truly independent within a subject. Further, the residual variability may differ across subjects and also within a subject over time. Indeed, Figure 3.1 makes clear that

subjects have different levels of variability in each activity metric. For these reasons, the direct use of Equation 3.1 and the prophecy formula may be an oversimplification for physical activity measurements. When applied to activity data, violations of these assumptions can produce inaccurate estimates of reliability R_1 and, by extension, $R_{\bar{J}}$.

Recall that to accurately assess the reliability R_1 of a measurement X, it is necessary to obtain several independent observations across multiple subjects. A similar argument applies to the reliability $R_{\bar{J}}$ of an aggregate measure $\bar{X}$ – independent replicates of the aggregate measure should be used to establish $R_{\bar{J}}$. A difference between the value obtained this way and one suggested by the prophecy formula may suggest that the prophecy formula used a flawed estimate of R_1. To our knowledge, however, a direct examination of $R_{\bar{J}}$ has not been conducted.

3.3 Methods

We now address two specific aspects of measurement reliability in the physical activity setting. First, we look to better understand the estimation of reliability and the use of the prophecy formula when replicate measurements are observed over a single period. Second, we assess the reliability of the average of several observations by taking such averages in two distinct observation windows. The first is intended to provide more insight into how reliability has been estimated in existing work based on single observation periods, and the second is intended to clarify the relationship between prophesied and actual reliability. Through addressing these specific aspects, we intend to inform how many valid days of wear are needed for the measurement of physical activity metrics via wearable devices.

For the first goal, we are interested especially in the variability of estimates of both R_1 and the number of days needed to obtain an aggregate reliability of 0.8. We quantify this sampling variability by randomly selecting an observation window for each subject of J days and calculating the single-measurement reliability $\hat{R}_1$ associated with sedentary, LPA, and MVPA time. Given the estimated reliability $\hat{R}_1$, we then determine the number of days $\hat{J}_{0.8}$ which would need to be averaged to achieve an aggregate reliability $R_{\bar{J}} = 0.8$ based on the the prophecy formula. This

process mimics the estimation of reliability and the application of the prophecy formula in a study with a single observation window. We repeat these steps 200 times each for $J = 2, 3, \ldots, 10$ valid days in order to obtain a sampling distribution of $\hat{R}_1$ and $\hat{J}_{0.8}$. Figure 3.2 contains a conceptual illustration of this process.

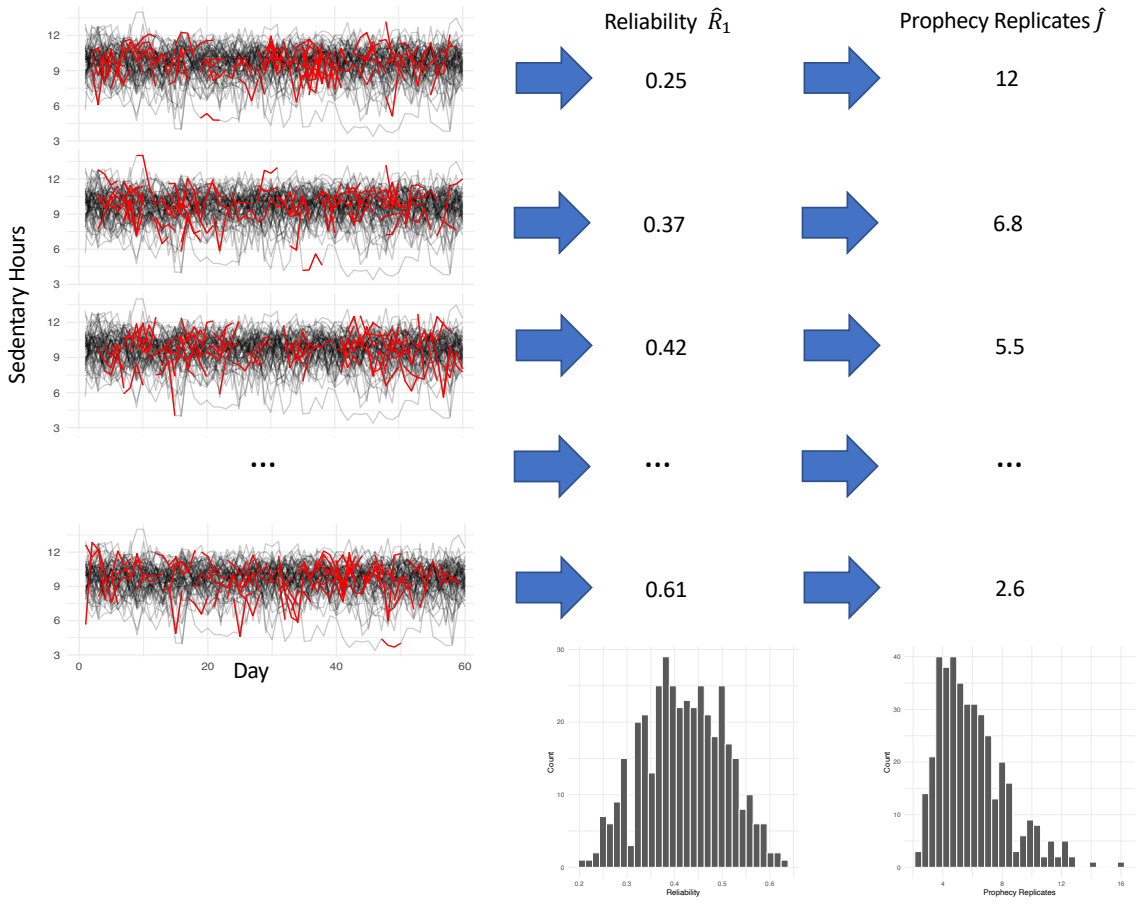


Figure 3.2: Framework for creating sampled datasets. Each row represents a single sampled dataset. The left column contains all of the observed activity trajectories in the motivating data, with the red segments representing the selection of days J which were selected for a given sample. The second column shows the estimated reliability $\hat{R}_1$ based on the previous selection of days and the final column the subsequent prophecy estimate of the number of replicates $\hat{J}$ required to achieve a reliability of 0.8. The histograms at the bottom of the second and third columns represent the distributions of $\hat{R}_1$ and $\hat{J}$ across all 200 samples.

Next, we look to determine the actual reliability for the averages of J valid days. We investigate this by first selecting two distinct periods of J days per subject and averaging sedentary, LPA, and MVPA time within both periods. We additionally require that the two periods be at least

7 days apart; this reduces potential within-subject correlation between the two distinct periods, although days within periods may still be correlated. In this way, we obtain two independent average measures of each activity metric, with averages based on observation windows of J days. Using these averages we then estimate $\hat{R}_{\bar{J}}$ directly through mixed model analysis based on the definition of reliability. This process is repeated 200 times each for $J = 2, 3, \ldots, 10$ days.

3.4 Results

3.4.1 Estimate Variability

Results for our first collection of analyses are shown in Figure 3.3. The top row shows the empirical distribution of estimated reliability for sedentary, LPA, and MVPA time in left, middle, and right panels respectively. Each panel shows the distribution of reliability estimates $\hat{R}_1$ based on observation periods of between 2 and 10 days with individual points showing the results for a single generated dataset. The bottom row of Figure 3.3 shows the corresponding empirical distribution of the number of replicates $\hat{J}_{0.8}$ suggested by the prophecy formula in order to achieve $R_{\bar{J}} = 0.8$ for each activity metric and observation window.

The results in Figure 3.3 emphasize that there is considerable variability in the estimates of R_1 and J, especially for lower numbers of observation days. As expected, with a larger number of days we see decreased variability in the estimation of both values. Across the activity metrics, reliability is highest for LPA time; correspondingly, the estimated number of replicates $\hat{J}$ obtained by the prophecy formula is smallest for LPA time. For example, the top center panel of Figure 3.3 shows that given 7 days of observation per subject, the median reliability estimate for LPA time was 0.51 (IQR = [0.47, 0.54]), with a minimum observed reliability of 0.40 and a maximum of 0.64. The bottom center panel then shows that given 7 days of observation per subject, the median number of replicates obtained by the prophecy formula was 3.8 (IQR = [3.4, 4.5]) with a minimum of 2.3 and maximum of 5.9. In contrast, the reliability $\hat{R}_1$ for MVPA obtained from a single sampled dataset is often lower than 0.4, and values for $\hat{J}_{0.8}$ are higher than 6. The values for R_1 and J obtained across sampled datasets is more variable for MVPA than for LPA or sedentary

time, which may derive from the heterogeneity within and across subjects seen in Figure 3.1.

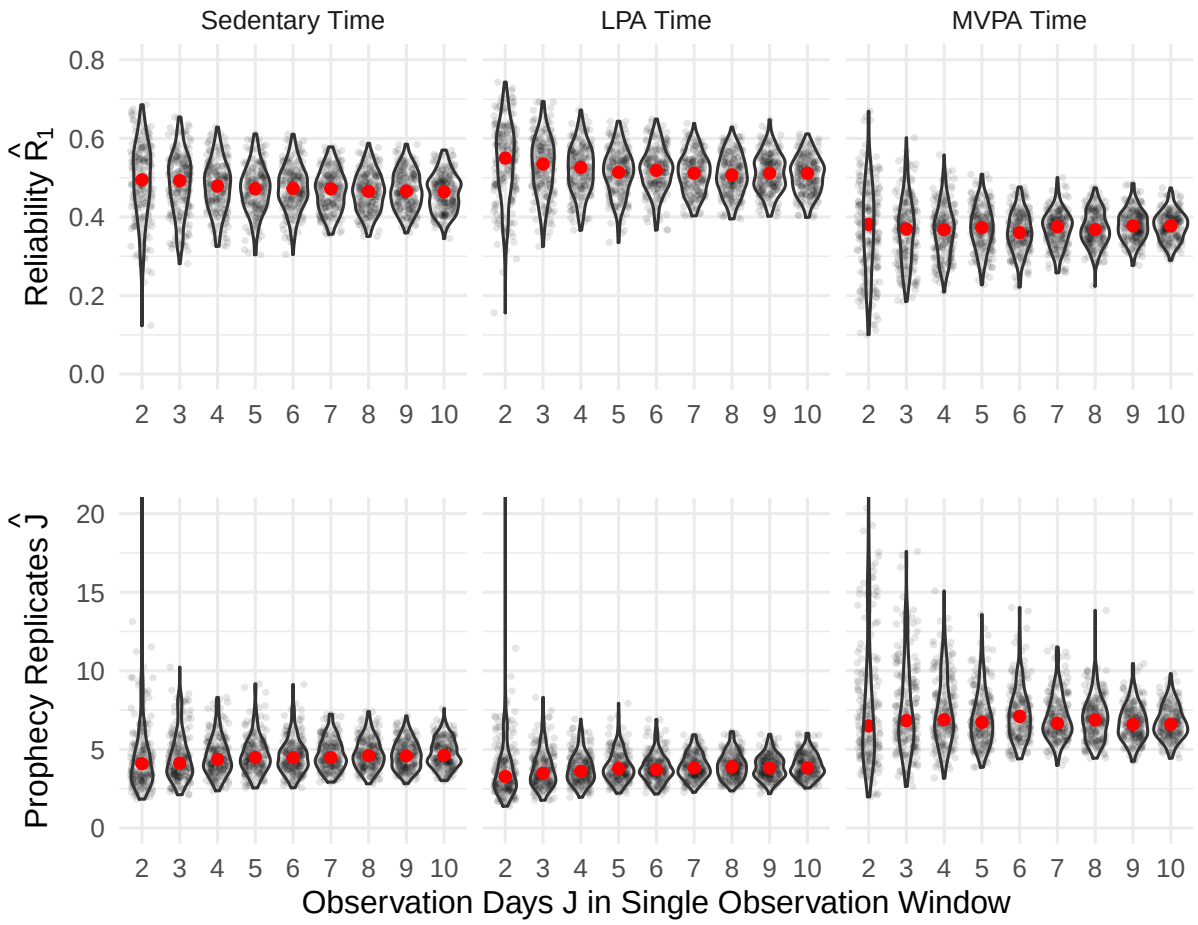


Figure 3.3: Results of the first investigation estimating R_1 and J based on a varying number of observation days. The top and bottom row show the distribution of reliability estimates $\hat{R}_1$ and number of days of observation suggested by the prophecy formula $\hat{J}$ for each activity metric, respectively, with the red point indicating the median across sampled datasets.

3.4.2 Aggregate Reliability Estimation

Figure 3.4 provides the results of our second analysis, in which two separate periods of J days were averaged for each subject to directly estimate aggregate reliability $\hat{R}_{\bar{J}}$. Results for sedentary,

LPA, and MVPA time are shown in the left, middle, and right panels respectively.

As expected, the aggregate reliability $\hat{R}_{\bar{J}}$ increased as the number of observation days J increased and the variability in $\hat{R}_{\bar{J}}$ decreased. Aggregate reliability was highest among LPA time, with the center panel of Figure 3.4 showing that for 7 days of observation the median $\hat{R}_{\bar{J}}$ was 0.83 (IQR = [0.80, 0.85]), with a minimum of 0.72 and a maximum of 0.89, with 77% of $\hat{R}_{\bar{J}} \geq 0.8$. In contrast to LPA, aggregate reliability was lowest within MVPA with a median $\hat{R}_{\bar{J}}$ of 0.75 (IQR = [0.71, 0.79]) for 7 days of observation, and only 22.5% of $\hat{R}_{\bar{J}} \geq 0.8$. These results also suggest that 10 valid observation days are needed to achieve a reliability of 0.8 for MVPA.

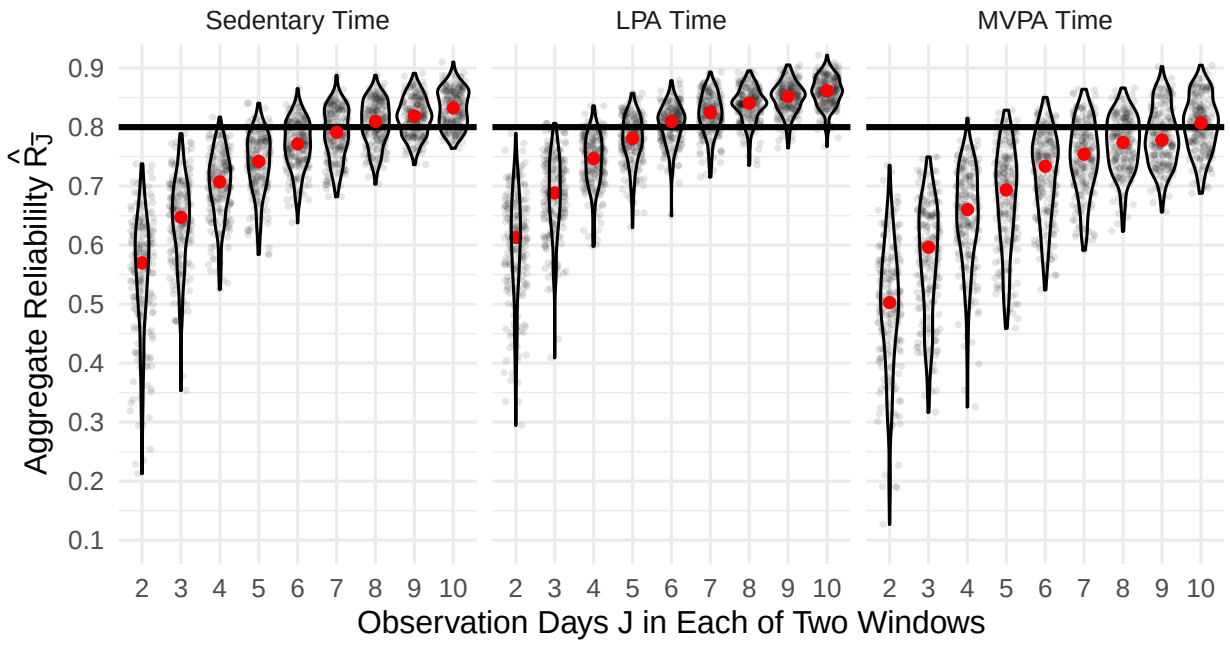


Figure 3.4: Results of the investigation into aggregate reliability for each activity metric, with two separate periods of between 2 and 10 day selected for each subject and averaged, with the subsequent reliability determined. The red point indicates the median across sampled datasets.

3.5 Discussion

Studies that use wearable devices to assess physical activity face a number of constraints that limit the number of observation days that can be gathered for each participant. In order to design these studies, researchers seek to collect sufficient data to ensure that resulting aggregate measurements are reliable for participants' habitual activity. Current recommendations were established using the prophecy formula applied to estimated reliability values from single observation windows.

Our work had two primary goals, which were made possible through a long-term study of physical activity. First, we sought to use the long-term nature of our data to better understand the properties of reliability and the prophecy formula as they have been used in the past. Second, we sought to assess aggregate reliability directly, through the use of multiple independent aggregate measurements made on each subject.

The results of our first investigation are broadly consistent with previous recommendations. That is, given sufficient pilot data, it would appear reasonable to conclude that 3-5 days of observation can be adequate to provide an aggregate reliability of 0.8 for sedentary and LPA time, and that 6-8 days of observation can be reasonable for MVPA time. Our work gives additional insight into the uncertainty in estimating reliability and the number of days necessary to obtain an aggregated reliability in a future study. Our results also highlight that these conclusions are dependent on the activity metric of interest.

The results of our second investigation, however, may suggest that the current approach to assessing reliability is flawed. In particular, the reliability of an aggregate measure based on 3-5 valid days is notably lower than 0.8. For sedentary time, 8 days were required in order to achieve a median reliability ≥ 0.8 across our sampled datasets. Similarly, 6 days were needed for LPA time, and 10 days for MVPA time in order to achieve a median reliability ≥ 0.8. These results indicate that a 7-day observation protocol is inadequate for sedentary and MVPA time, particularly when allowing for non-compliance or invalid observation days. Longer studies and stricter adherence

may both be necessary to achieve an expected reliability ≥ 0.8 across physical activity metrics.

The contrast between results in our first and second investigations may not be as surprising as they initially appear. The reliability framework developed within psychometrics, where the assumptions of the data generating model in Equation 3.1 may be valid. Extending this framework to sequential measurements of activity over time is imperfect due to a lack of independence within subject measurements, and non-constant variance across subjects. When assumptions are not met in practice, estimates of reliability based on single observation windows may be biased. By constructing aggregate measures in two distinct time windows, we are able to obtain independent averages and assess reliability directly.

Our results stem from the analysis of a single long-term followup cohort on a specific population and are not guaranteed to apply to other cohorts. Although we took efforts to ensure that the data we considered would not contain changes in participant level habitual activity, it's possible that such changes are present and affect our results. We suspect that issues of within-subject correlation and non-constant variance across subjects are the main drivers behind the gap between prophesied and real aggregate reliability. Analysis tools that account for these might improve estimation of R_1 and J in data based on a single observation window and help close the observed gap. Our results are derived from a sample of 69 subjects, and some of the sampling variability we observe in Figures 3.3 and 3.4 would be reduced in analyses of larger cohorts. That said, the median values of R_1, J, and $R_{\bar{J}}$ and our conclusions about the duration of follow-up are not likely to be affected by larger sample sizes.

This manuscript is focused on better understanding the framework that has been used for assessing the reliability of sedentary, LPA, and MVPA time. Our results suggest a mismatch between the assumed data generating structure in Equation 3.1 and the values produced in studies of physical activity. These may be addressed through improved analysis methods, but other critiques of the general approach will remain. Reliability is a difficult quantity to interpret. When data are indeed generated under Equation 3.1, R_1 is identical to the Pearson correlation between observations made at two time points [53]; the definition of R_1 is also recognizable as the intraclass correlation coeffi-

cient (ICC) used in mixed models. While this could be helpful in building intuition for reliability, it may serve instead to confuse: in general Pearson correlations and the ICC are not the same. Further, a definition of reliability that depends on the ratio of between subject to total variability will mean that the same measurement could be more or less reliable as the population in question changes. Finally, and related, reliability is distinct from accuracy, and even reliable measurements may not accurately reflect the underlying quantity in question. Whether these criticisms require an alternative framework for understanding reliability is, as yet, unclear.

3.6 Conclusions

Our results suggest that while it is common to estimate a number of monitoring days between 3-5 and 6-8 for sedentary/LPA and MVPA time, respectively, in practice, the reliability of measurements using a monitoring period of the previous size more often than not does not result in an aggregate reliability of 0.8. In practice, a protocol of at least 8 days of observation is needed to assert that 0.8 reliability is likely to be met in terms of sedentary time, with protocol lengths of 6 and 10 days for LPA and MVPA time respectively. Significantly longer protocol lengths should be considered to definitively assert that a reliability of 0.8 has been achieved.

Conclusion

This thesis has sought to address a number of pressing issues within wearable device research including; what is to be done when individuals are not compliant with device wear during the study, how large outlying values can be identified and adjusted for, and finally how many observation days are necessary in order to reliably estimate the average daily activity metrics of individuals under study.

The results of Chapter 1 have shown that in the presence of missing data, an approach which accounts for the time which missing data was observed, as well as a subject's observed activity during periods of wear, can provide a more accurate estimate of activity during nonwear periods than approaches which impute a single average activity value for all subjects and time points. Within Chapter 2 we provided a novel approach for identifying and handling large outlying values which leveraged existing methods in FPCA and PCP, and showed that an observed activity profile $y_i(t)$ can be segmented into low-rank, sparse, and random error components, providing a more accurate estimate of underlying activity than a standard application of FPCA, and additionally allowing for identifying the location of sparse outliers which may themselves be of interest in many situations. Finally, in Chapter 3 we presented an in-depth analysis of measurement reliability using a long-term followup cohort of activity measurement data. The results of this analysis shed light onto the variability in which single measurement reliability, and subsequently the suggested number of days to be used in a future study, is estimated. Further, using the long-term nature of our cohort we showed that practical reliability is often less than that prophesied by the Spearman-Brown formula, and note the assumptions of the prophecy formula which are often not applicable

in the physical activity setting and contribute to this discrepancy.

References

[1] R. P. Troiano, J. J. McClain, R. J. Brychta, and K. Y. Chen, "Evolution of accelerometer methods for physical activity research," Br J Sports Med, bjsports–2014, 2014.

[2] W. Willett and M. J. Stampfer, "Total energy intake: Implications for epidemiologic analyses.," American journal of epidemiology, vol. 124, no. 1, pp. 17–27, 1986.

[3] G. N. Healy, C. E. Matthews, D. W. Dunstan, E. A. Winkler, and N. Owen, "Sedentary time and cardio-metabolic biomarkers in US adults: NHANES 2003–06," European heart journal, vol. 32, no. 5, pp. 590–597, 2011.

[4] Q. Qi, G. Strizich, G. Merchant, D. Sotres-Alvarez, C. Buelna, S. F. Castañeda, L. C. Gallo, J. Cai, M. D. Gellman, C. R. Isasi, et al., "Objectively measured sedentary time and cardiometabolic biomarkers in US Hispanic/Latino adults: the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)," Circulation, vol. 132, no. 16, pp. 1560–1569, 2015.

[5] K. R. Evenson, F. Wen, and A. H. Herring, "Associations of accelerometry-assessed and self-reported physical activity and sedentary behavior with all-cause and cardiovascular mortality among us adults," American journal of epidemiology, vol. 184, no. 9, pp. 621–632, 2016.

[6] B. J. Jefferis, T. J. Parsons, C. Sartini, S. Ash, L. T. Lennon, O. Papacosta, R. W. Morris, S. G. Wannamethee, I.-M. Lee, and P. H. Whincup, "Objectively measured physical activity, sedentary behaviour and all-cause mortality in older men: Does volume of activity matter more than pattern of accumulation?" British journal of sports medicine, vol. 53, no. 16, pp. 1013–1020, 2019.

[7] I.-M. Lee, E. J. Shiroma, M. Kamada, D. R. Bassett, C. E. Matthews, and J. E. Buring, "Association of step volume and intensity with all-cause mortality in older women," JAMA internal medicine, vol. 179, no. 8, pp. 1105–1112, 2019.

[8] P. C. Zee, H. Attarian, and A. Videnovic, "Circadian rhythm abnormalities," Continuum: Lifelong Learning in Neurology, vol. 19, no. 1 Sleep Disorders, p. 132, 2013.

[9] J. O. Ramsay, Functional data analysis. Wiley Online Library, 2006.

[10] P. Kokoszka and M. Reimherr, Introduction to functional data analysis. CRC Press, 2017.

[11] H. Sørensen, J. Goldsmith, and L. M. Sangalli, "An introduction with medical applications to functional data analysis," Statistics in medicine, vol. 32, no. 30, pp. 5222–5240, 2013.

[12] M Cortina-Borja, M Geraci, L Griffiths, C Rich, C Dezateux, et al., "Modelling accelerometer data from 7-year old british children using functional analysis of variance," J Epidemiol Community Health, vol. 65, no. Suppl 2, A26–A27, 2011.

[13] H. Shou, V. Zipunnikov, C. M. Crainiceanu, and S. Greven, "Structured functional principal component analysis," Biometrics, vol. 71, no. 1, pp. 247–257, 2015.

[14] J. Goldsmith, V. Zipunnikov, and J. Schrack, "Generalized multilevel function-on-scalar regression and principal component analysis," Biometrics, vol. 71, no. 2, pp. 344–353, 2015.

[15] J. Goldsmith, X. Liu, J. Jacobson, and A. Rundle, "New insights into activity patterns in children, found using functional data analyses," Medicine and science in sports and exercise, vol. 48, no. 9, p. 1723, 2016.

[16] M. M. Burg, J. E. Schwartz, I. M. Kronish, K. M. Diaz, C. Alcantara, J. Duer-Hefele, and K. W. Davidson, "Does stress result in you exercising less? or does exercising result in you being less stressed? or is it both? testing the bi-directional stress-exercise association at the group and person (n of 1) level," Annals of Behavioral Medicine, pp. 1–11, 2017.

[17] S. Yoon, J. E. Schwartz, M. M. Burg, I. M. Kronish, C. Alcantara, J. Julian, F. Parsons, K. W. Davidson, and K. M. Diaz, "Using behavioral analytics to increase exercise: A randomized n-of-1 study," American journal of preventive medicine, vol. 54, no. 4, pp. 559–567, 2018.

[18] B. Hutto, V. J. Howard, S. N. Blair, N. Colabianchi, J. E. Vena, D. Rhodes, and S. P. Hooker, "Identifying accelerometer nonwear and wear time in older adults," International Journal of Behavioral Nutrition and Physical Activity, vol. 10, no. 1, p. 120, 2013.

[19] S. G. Trost, K. L. McIver, and R. R. Pate, "Conducting accelerometer-based activity assessments in field-based research," Medicine & Science in Sports & Exercise, vol. 37, no. 11, S531–S543, 2005.

[20] F. Yao, H.-G. Müller, and J.-L. Wang, "Functional data analysis for sparse longitudinal data," Journal of the American Statistical Association, vol. 100, no. 470, pp. 577–590, 2005.

[21] J. Goldsmith, S. Greven, and C. Crainiceanu, "Corrected confidence bands for functional data using principal components," Biometrics, vol. 69, no. 1, pp. 41–51, 2013.

[22] I. Jolliffe, Principal Component Analysis. Springer, 1986.

[23] P. Hall, H.-G. Müller, and F. Yao, "Modelling sparse generalized longitudinal observations with latent gaussian processes,"

Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 70, no. 4, pp. 703–723, 2008.

[24] J. Gertheiss, J. Goldsmith, and A.-M. Staicu, "A note on modeling sparse exponential-family functional response curves," Computational statistics & data analysis, vol. 105, pp. 46–52, 2017.

[25] D. Backenroth, R. T. Shinohara, J. A. Schrack, and J. Goldsmith, "Nonnegative decomposition of functional count data," Biometrics, 2020.

[26] B. Efron and R. J. Tibshirani, An introduction to the bootstrap. CRC press, 1994.

[27] J. Wrobel, V. Zipunnikov, J. Schrack, and J. Goldsmith, "Registration for exponential family functional data," Biometrics, vol. 75, no. 1, pp. 48–57, 2019.

[28] L. Huang, J. Bai, A. Ivanescu, T. Harris, M. Maurer, P. Green, and V. Zipunnikov, "Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations," Journal of the American Statistical Association, vol. 114, no. 526, pp. 553–564, 2019.

[29] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" Journal of the ACM (JACM), vol. 58, no. 3, pp. 1–37, 2011.

[30] J. Quesada and P. Rodriguez, "Automatic vehicle counting method based on principal component pursuit background modeling," in 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 3822–3826.

[31] Y. Pan, C. Yang, R. An, and Y. Sun, "Robust principal component pursuit for fault detection in a blast furnace process," Industrial & Engineering Chemistry Research, vol. 57, no. 1, pp. 283–291, 2018.

[32] T. Bouwmans and E. H. Zahzah, "Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance," Computer Vision and Image Understanding, vol. 122, pp. 22–34, 2014.

[33] Z. Yan, C.-Y. Chen, L. Luo, and Y. Yao, "Stable principal component pursuit-based thermographic data analysis for defect detection in polymer composites," Journal of Process Control, vol. 49, pp. 36–44, 2017.

[34] K. Min, Z. Zhang, J. Wright, and Y. Ma, "Decomposing background topics from keywords by principal component pursuit," in Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 269–278.

[35] G. M. James, "Generalized linear models with functional predictors," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 64, no. 3, pp. 411–432, 2002.

[36] A. Van Der Linde, "A bayesian latent variable approach to functional principal components analysis with binary and count data," AStA Advances in Statistical Analysis, vol. 93, no. 3, pp. 307–333, 2009.

[37] L. Xiao, V. Zipunnikov, D. Ruppert, and C. Crainiceanu, "Fast covariance estimation for high-dimensional functional data," Statistics and computing, vol. 26, no. 1-2, pp. 409–421, 2016.

[38] E. Grave, G. R. Obozinski, and F. Bach, "Trace lasso: A trace norm regularization for correlated designs," Advances in Neural Information Processing Systems, vol. 24, pp. 2187–2195, 2011.

[39] X. Yuan and J. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," preprint, vol. 12, no. 2, 2009.

[40] J. Goldsmith, F. Scheipl, L. Huang, J. Wrobel, C. Di, J. Gellar, J. Harezlak, M. W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, and P. T. Reiss, Refund: Regression with functional data, R package version 0.1-23, 2020.

[41] L. R. Dice, "Measures of the amount of ecologic association between species," Ecology, vol. 26, no. 3, pp. 297–302, 1945.

[42] J. Bai, B. He, H. Shou, V. Zipunnikov, T. A. Glass, and C. M. Crainiceanu, "Normalization and extraction of interpretable metrics from raw accelerometry data," Biostatistics, vol. 15, no. 1, pp. 102–116, 2014.

[43] S. Levin, D. R. Jacobs Jr, B. E. Ainsworth, M. T. Richardson, and A. S. Leon, "Intra-individual variation and estimates of usual physical activity," Annals of epidemiology, vol. 9, no. 8, pp. 481–488, 1999.

[44] S. G. Trost, R. R. Pate, P. S. Freedson, J. F. Sallis, and W. C. Taylor, "Using objective physical activity measures with youth: How many days of monitoring are needed?" Medicine & Science in Sports & Exercise, vol. 32, no. 2, p. 426, 2000.

[45] C. E. Matthews, B. E. Ainsworth, R. W. Thompson, and D. R. Bassett Jr, "Sources of variance in daily physical activity levels as measured by an accelerometer.," Medicine and science in sports and exercise, vol. 34, no. 8, pp. 1376–1381, 2002.

[46] M. S. Treuth, N. E. Sherwood, N. F. Butte, B. McClanahan, E. Obarzanek, A. Zhou, C. Ayers, A. Adolph, J. Jordan, D. R. Jacobs, et al., "Validity and reliability of activity

measures in african-american girls for gems.," Medicine and science in sports and exercise, vol. 35, no. 3, pp. 532–539, 2003.

[47] D. M. Murray, J. Stevens, P. J. Hannan, D. J. Catellier, K. H. Schmitz, M. Dowda, T. L. Conway, J. C. Rice, and S. Yang, "School-level intraclass correlation for physical activity in sixth grade girls," Medicine and science in sports and exercise, vol. 38, no. 5, p. 926, 2006.

[48] N. Mayerson and R. Milano, "Goniometric measurement reliability in physical medicine.," Archives of physical Medicine and Rehabilitation, vol. 65, no. 2, pp. 92–94, 1984.

[49] M. L. Leff, J. O. Hill, A. A. Yates, G. A. Cotsonis, and S. B. Heymsfield, "Resting metabolic rate: Measurement reliability," Journal of Parenteral and enteral Nutrition, vol. 11, no. 4, pp. 354–359, 1987.

[50] W. Talloen, S. Hochreiter, L. Bijnens, A. Kasim, Z. Shkedy, D. Amaratunga, and H. Göhlmann, "Filtering data from high-throughput experiments based on measurement reliability," Proceedings of the National Academy of Sciences, vol. 107, no. 46, E173–E174, 2010.

[51] K. L. Pearce, A. Sufrinko, B. C. Lau, L. Henry, M. W. Collins, and A. P. Kontos, "Near point of convergence after a sport-related concussion: Measurement reliability and relationship to neurocognitive impairment and symptoms," The American journal of sports medicine, vol. 43, no. 12, pp. 3055–3061, 2015.

[52] X.-N. Zuo, T. Xu, and M. P. Milham, "Harnessing reliability for neuroscience research," Nature human behaviour, vol. 3, no. 8, pp. 768–771, 2019.

[53] J. L. Fleiss, Design and analysis of clinical experiments. John Wiley & Sons, 2011, vol. 73.

[54] D. S. Ward, K. R. Evenson, A. Vaughn, A. B. Rodgers, and R. P. Troiano, "Accelerometer use in physical activity: Best practices and research recommendations.," Medicine and science in sports and exercise, vol. 37, no. 11 Suppl, S582–8, 2005.

[55] R. P. Troiano, D. Berrigan, K. W. Dodd, L. C. Mâsse, T. Tilert, M. McDowell, et al., "Physical activity in the united states measured by accelerometer," Medicine and science in sports and exercise, vol. 40, no. 1, p. 181, 2008.

[56] N. L. Glazer, A. Lyass, D. W. Esliger, S. J. Blease, P. S. Freedson, J. M. Massaro, J. M. Murabito, and R. S. Vasan, "Sustained and shorter bouts of physical activity are related to cardiovascular health," Medicine and science in sports and exercise, vol. 45, no. 1, p. 109, 2013.

[57] K. R. Evenson, D. Sotres-Alvarez, Y. Deng, S. J. Marshall, C. R. Isasi, D. W. Esliger, and S. Davis, "Accelerometer adherence and performance in a cohort study of us hispanic adults," Medicine and science in sports and exercise, vol. 47, no. 4, p. 725, 2015.

[58] V. J. Howard, J. D. Rhodes, A. Mosher, B. Hutto, M. S. Stewart, N. Colabianchi, J. E. Vena, S. N. Blair, and S. P. Hooker, "Obtaining accelerometer data in a national cohort of black and white adults," Medicine and science in sports and exercise, vol. 47, no. 7, p. 1531, 2015.

[59] I.-M. Lee, E. J. Shiroma, K. R. Evenson, M. Kamada, A. Z. LaCroix, and J. E. Buring, "Using devices to assess physical activity and sedentary behavior in a large cohort study: The women's health study," Journal for the measurement of physical behaviour, vol. 1, no. 2, pp. 60–69, 2018.

[60] S Baumann, S Groß, L Voigt, A Ullrich, F Weymar, T Schwaneberg, M Dörr, C Meyer, U John, and S Ulbricht, "Pitfalls in accelerometer-based measurement of physical activity: The presence of reactivity in an adult population," Scandinavian journal of medicine & science in sports, vol. 28, no. 3, pp. 1056–1063, 2018.

[61] C. Tudor-Locke, E. J. Aguiar, H. Han, S. W. Ducharme, J. M. Schuna, T. V. Barreira, C. C. Moore, M. A. Busa, J. Lim, J. R. Sirard, et al., "Walking cadence (steps/min) and intensity in 21–40 year olds: Cadence-adults," International Journal of Behavioral Nutrition and Physical Activity, vol. 16, no. 1, pp. 1–11, 2019.

[62] C. Tudor-Locke, S. W. Ducharme, E. J. Aguiar, J. M. Schuna, T. V. Barreira, C. C. Moore, C. J. Chase, Z. R. Gould, M. A. Amalbert-Birriel, J. Mora-Gonzalez, et al., "Walking cadence (steps/min) and intensity in 41 to 60-year-old adults: The cadence-adults study," International Journal of Behavioral Nutrition and Physical Activity, vol. 17, no. 1, pp. 1–10, 2020.

Appendix A: Similarities in Adjustment Approaches

In this appendix we discuss the similarities between adjusting for corrected steps, using the linear model approach, and adjusting for both observed activity and weartime when the goal is to assess the effect of activity on an outcome of interest. Within the **Analytic Results** Section we show that adjusting for corrected steps using the linear model approach is equivalent to fitting a model with both observed activity and weartime, given an often reasonable constraint on the relationship between weartime and the outcome. Additionally, we show that when no additional covariates are included in the respective models, the coefficients associated with activity will always be identical for these approaches, irrespective of the constraint. In the presence of additional covariates, a closed form solution is not easily derived, and as such within the **Simulation Results** Section we generate simulated datasets with a known coefficient associated with activity, and illustrate that both approaches have similar characteristics and accuracy in recovering the known activity coefficient under a number of different simulation scenarios.

A.1 Analytic Results

A.1.1 Derivation of Constrained Regression

Below we show that adjusting for the corrected steps, via the linear model approach, is equivalent to a constrained regression version of the model which adjusts for both observed activity and weartime.

The linear model weartime adjustment is derived from the model

$$\text{activity}_{obs} = \gamma_0 + \gamma_1 \cdot \text{weartime}_{obs} + \epsilon$$

where activity_{obs} and weartime_{obs} are the observed activity and weartime, and ϵ is a mean zero error term. When applying the linear model correction, the corrected activity $\left(\text{activity}_{cor}\right)$ is given by

$$\text{activity}_{cor} = \text{activity}_{obs} + (\text{weartime}_{max} - \text{weartime}_{obs}) \cdot \gamma_1$$

where weartime_{max} is the maximum observable weartime.

When we are interested in assessing the effect of activity on an outcome Y, we consider the following models

$$Y = \beta_0 + \beta_1 \cdot \text{activity}_{cor} + \epsilon \tag{A.1}$$

$$Y = \alpha_0 + \alpha_1 \cdot \text{activity}_{obs} + \alpha_2 \cdot \text{weartime}_{obs} + \epsilon \tag{A.2}$$

From model A.1

$$\begin{aligned} Y &= \beta_0 + \beta_1 \cdot \text{activity}_{cor} + \epsilon \\ &= \beta_0 + \beta_1 \cdot \left(\text{activity}_{obs} + (\text{weartime}_{max} - \text{weartime}_{obs}) \cdot \gamma_1\right) + \epsilon \\ &= \beta_0 + \beta_1 \cdot \text{activity}_{obs} + \beta_1 \cdot \text{weartime}_{max} \cdot \gamma_1 - \beta_1 \cdot \text{weartime}_{obs} \cdot \gamma_1 + \epsilon \end{aligned}$$

Combining the above, model A.1 represents a constrained version of model A.2, with the constraint being

$$\alpha_2 = -\alpha_1 \cdot \gamma_1$$

This constraint multiplies the coefficient for activity on the outcome by the coefficient for weartime on activity – roughly, this is the expected increase in activity for a unit increase in

weartime multiplied by the expected change in the outcome for each unit change in activity. The above result is invariant to the inclusion of additional covariates in models A.1 and A.2.

A.1.2 Equivalence in the Absence of Other Covariates

Below we show that when no additional covariates are included in either model, then adjusting for the corrected steps or observed activity and weartime will always result in identical coefficients for the activity covariates in both models.

It is known that for model A.1

$$\beta_1 = \frac{cov(Y, \text{activity}_{cor})}{var(\text{activity}_{cor})}$$

and similarly within the LM adjustment model

$$\gamma_1 = \frac{cov(\text{activity}_{obs}, \text{weartime}_{obs})}{var(\text{weartime}_{obs})}$$

Additionally, for model A.2 it can be shown that

$$\alpha_1 = \frac{var(\text{weartime}_{obs}) \cdot cov(\text{activity}_{obs}, \text{Y}) - cov(\text{activity}_{obs}, \text{weartime}_{obs}) \cdot cov(\text{weartime}_{obs}, Y)}{var(\text{activity}_{obs}) \cdot var(\text{weartime}_{obs}) - cov(\text{activity}_{obs}, \text{weartime}_{obs})^2}$$

These results derive from known properties of linear regression models with one or two coefficients. Given the previous, when no additional covariates are included in models A.1 and A.2, the following holds

$$
\begin{aligned}
\beta_1 &= \frac{cov(Y, \text{activity}_{cor})}{var(\text{activity}_{cor})} \\
&= \frac{cov(Y, \text{activity}_{obs} + (\text{weartime}_{max} - \text{weartime}_{obs}) \cdot \gamma_1)}{var(\text{activity}_{obs} + (\text{weartime}_{max} - \text{weartime}_{obs}) \cdot \gamma_1)} \\
&= \frac{cov(Y, \text{activity}_{obs}) - \gamma_1 \cdot cov(Y, \text{weartime}_{obs})}{var(\text{activity}_{obs}) + \gamma_1^2 \cdot var(\text{weartime}_{obs}) - 2 \cdot \gamma_1 \cdot cov(\text{activity}_{obs}, \text{weartime}_{obs})} \\
&= \frac{cov(Y, \text{activity}_{obs}) - \gamma_1 \cdot cov(Y, \text{weartime}_{obs})}{var(\text{activity}_{obs}) + \gamma_1 \cdot (\gamma_1 \cdot var(\text{weartime}_{obs}) - 2 \cdot cov(\text{activity}_{obs}, \text{weartime}_{obs}))} \\
&= \frac{cov(Y, \text{activity}_{obs}) - \gamma_1 \cdot cov(Y, \text{weartime}_{obs})}{var(\text{activity}_{obs}) - \gamma_1 \cdot cov(\text{activity}_{obs}, \text{weartime}_{obs})}
\end{aligned}
$$

substituting the known value of γ_1 and rearranging terms, we have

$$
\begin{aligned}
&= \frac{cov(\text{Y}, \text{activity}_{obs}) - \frac{cov(\text{activity}_{obs}, \text{weartime}_{obs}) \cdot cov(Y, \text{weartime}_{obs})}{var(\text{weartime}_{obs})}}{var(\text{activity}_{obs}) - \frac{cov(\text{activity}_{obs}, \text{weartime}_{obs})^2}{var(\text{weartime}_{obs})}} \\
&= \frac{var(\text{weartime}_{obs}) \cdot cov(\text{activity}_{obs}, Y) - cov(\text{activity}_{obs}, \text{weartime}_{obs}) \cdot cov(\text{weartime}_{obs}, Y)}{var(\text{activity}_{obs}) var(\text{weartime}_{obs}) - cov(\text{activity}_{obs}, \text{weartime}_{obs})^2} \\
&= \alpha_1
\end{aligned}
$$

Thus without any additional covariates the regression coefficients associated with activity will always be identical between models A.1 and A.2.

A.2 Simulation Results

The previous has highlighted the equivalence of adjusting for the corrected steps, via the linear model approach, and adjusting for both observed activity and weartime in terms of the coefficient estimates for activity. In the presence of additional covariates, closed form solutions for regression coefficients are not readily available, and the previous derivations are prohibitive. As such we assess the differences in results using a simulation study.

We assess three specific scenarios, described in more detail below. Scenario 1 induces a low

correlation between observed activity and weartime while Scenario 2 induces a high correlation between observed activity and weartime. In both Scenarios 1 and 2, we adjust for a confounder that is associated with the outcome and weartime. Scenario 3 uses the same high correlation structure as Scenario 2, but omits the confounding covariate and examines the effect on coefficients for activity and weartime.

We generate 2000 simulated datasets comprised of 100 observations under each of the the previous scenarios. For each data set, we fit model A.1 which adjusts for the corrected steps and model A.2 which adjusts for observed activity and weartime. In both, we adjust for additional covariates, and investigate differences in the resulting coefficient estimates. For the activity coefficients, we assess the difference between the estimated value and the known value under the data generating process using the standardized difference $(\hat{\beta} - \beta)/\beta$. We also show the distribution of the weartime coefficient in order to assess differences across scenarios.

A.2.1 Simulation Design

Below we simulate data in the following way:

$$\text{weartime}_{obs} \sim unif(10, 14)$$

$$\text{activity}_{obs} = 5000 + 500 \cdot \text{weartime}_{obs} + \epsilon \text{ where } \epsilon \sim N\left(0, \sigma^2\right)$$

In Scenario 1, we set $\sigma^2 = 2000^2$ to induce low activity and weartime correlation, while in Scenario 2 and 3 we set $\sigma^2 = 500^2$ to induce high activity and weartime correlation.

activity_{cor} is calculated via the linear model correction as outlined previously

activity_{true} is the underlying activity which would have been observed given full weartime (i.e. what activity_{cor} attempts to estimate) and is given by

$\text{activity}_{true} = \text{activity}_{cor} + |\epsilon|$ where $\epsilon \sim N\left(0, 500^2\right)$

with $|\epsilon|$ taken to ensure that $\text{activity}_{true} \geq \text{activity}_{obs}$

cov_1 and cov_2 are additional covariates of interest where

$\text{cov}_1 = 50 + 4 \cdot \text{weartime}_{obs} + \epsilon$ where $\epsilon \sim N\left(0, 5^2\right)$ and is associated with weartime_{obs}

$\text{cov}_2 = 50 + \epsilon$ where $\epsilon \sim N\left(0, 5^2\right)$ and is independent of weartime_{obs}

The outcome Y is given by

$Y = 2 + 0.001 \cdot \text{activity}_{true} + 5 \cdot \text{cov}_1 + \epsilon$ where $\epsilon \sim N\left(0, 10^2\right)$

In the above cov_2 is not associated with the outcome Y

A.2.2 Average Correlation Matrices

Below we highlight the average correlation matrix across the simulations under each scenario

Scenario 1: Low Activity and Weartime Correlation

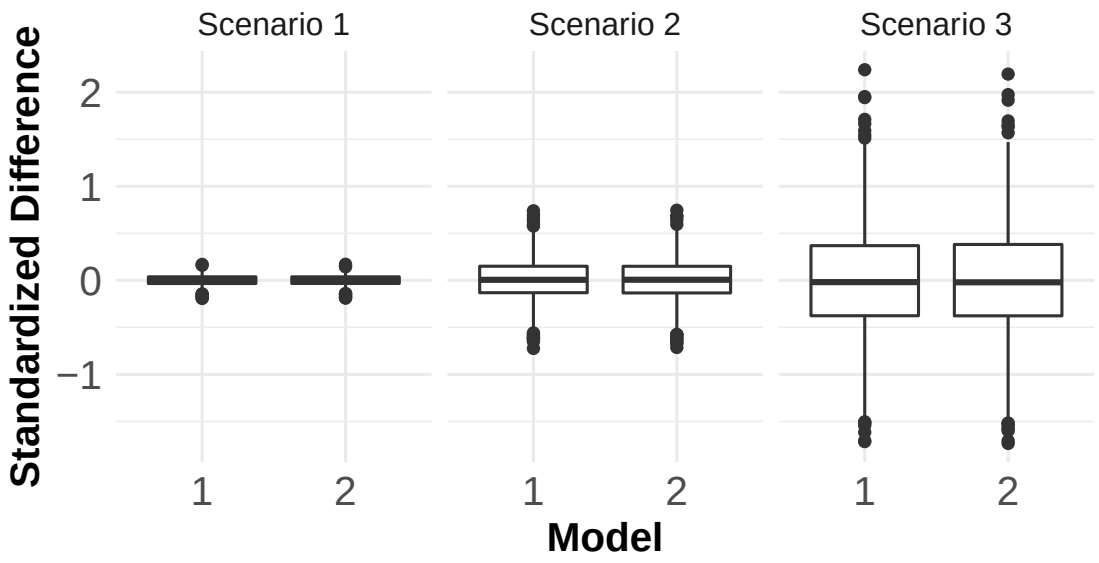
	Y	activity_obs	weartime	cov_1	cov_2
Y	1	0.61	0.65	0.86	0
activity_obs		1.00	0.28	0.21	0
weartime			1.00	0.75	0
cov_1				1.00	0
cov_2					1

Scenario 2 and 3: High Activity and Weartime Correlation

	Y	activity_obs	weartime	cov_1	cov_2
Y	1	0.63	0.72	0.96	0
activity_obs		1.00	0.76	0.57	0
weartime			1.00	0.76	0
cov_1				1.00	0
cov_2					1

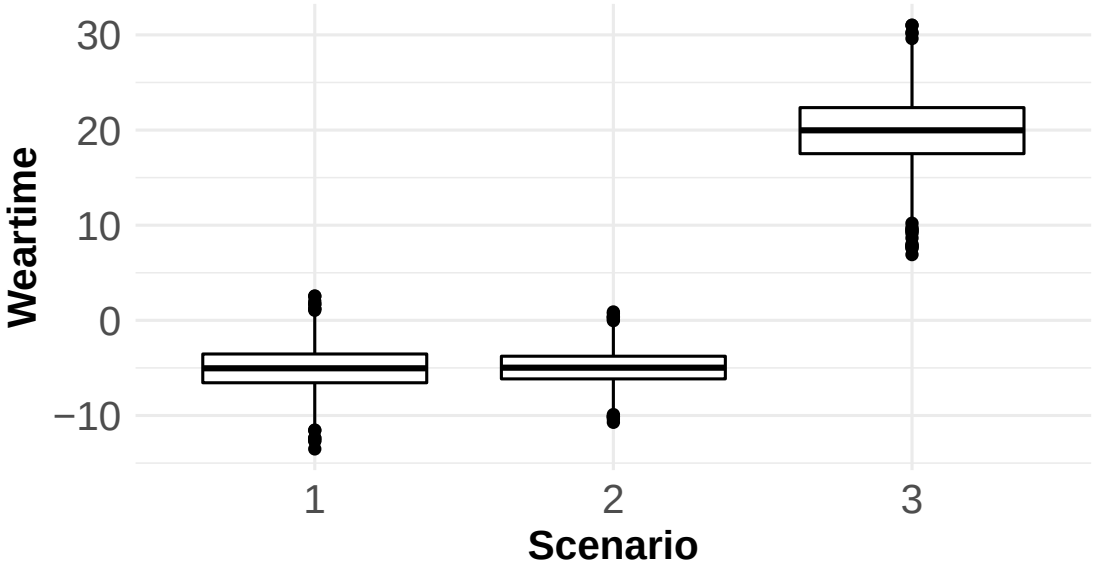
A.2.3 Activity Coefficient Estimates

The following figure shows the accuracy of each model in estimating the known coefficient associated with activity in the data generating process. Under all 3 scenarios, the accuracy of both models in estimating the known underlying activity covariate is similar. As the correlation increases between activity and weartime, we see larger standardized differences for both methods, and similarly when the confounding covariate is excluded in Scenario 3, we see larger standardized differences, but again similar between both methods. There are only small differences in the estimated coefficients for activity under either model.



A.2.4 Weartime Coefficient Estimates

Below we show the distribution of the weartime coefficient estimates from model A.2, adjusting for both observed activity and weartime. In Scenario 1 and 2 in which there is low and high correlation between activity and weartime, respectively, we see similar estimates of weartime, with some reduced variability in estimating in Scenario 2. In Scenario 3, in which a confounding covariate associated with both the outcome and weartime is excluded, we see large differences in the estimated coefficient associated with weartime, indicative of the omitted variable bias associated with the model under Scenario 3. Note that this affects the coefficient for weartime, but does not impact the coefficient for activity.


Weartime
30
20
10
0
−10
1
2
3
Scenario