STATISTICAL METHODS IN APPLICATIONS WITH COMPLEX DATA STRUCTURES

by

Jordan Johns

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

May, 2019

© 2019 Jordan Johns
All rights reserved
Abstract

This thesis details three applications of statistical methods, motivated by data sets with complex data structures and unique problems. In each case, the data structure and complications are described, followed by a detailed account of the approach taken to account for these complications and the corresponding conclusions concerning each methods’ feasibility regarding how each method can be applied in other scenarios. The three motivating examples are: 1) Dynamic prediction of MRI intensities in a longitudinal study of individuals with multiple sclerosis (MS), 2) Fragmentation in normalized trajectories of mood and attention measured by ecological momentary assessments (EMA), and 3) Variable-domain functional principal component analysis for short-term mortality score trajectories in individuals with acute respiratory distress in a hospital intensive care unit (ICU).
Thesis Committee

Primary Readers

Ciprian Crainiceanu (Primary Advisor)
    Professor
    Department of Biostatistics
    Johns Hopkins Bloomberg School of Public Health

Vadim Zipunnikov
    Associate Professor
    Department of Biostatistics
    Johns Hopkins Bloomberg School of Public Health

Adam Spira
    Associate Professor
    Department of Mental Health
    Johns Hopkins Bloomberg School of Public Health

Jennifer Schrack
    Associate Professor
    Department of Epidemiology
    Johns Hopkins Bloomberg School of Public Health

Kathleen Merikangas
    Adjunct Professor
    Departments of Mental Health and Epidemiology
    Johns Hopkins Bloomberg School of Public Health
Alternate Readers

Martin Lindquist
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Sapna Kudchadkar
Associate Professor
Department of Anesthesiology and Critical Care Medicine
Johns Hopkins School of Medicine
Acknowledgments

I would first like to thank my wonderful family, particularly my amazing wife Kimberlee, and my two children, Brooklyn and Matthew. This work would not be possible without your support and love, and the belief that I could do it. Also, to my parents for their example growing up and helping me understand how to put the important things in life first.

I would also like to thank the Biostatistics Department here at Johns Hopkins for the help, support, and camaraderie I have felt while I have been here. They have done an wonderful job encouraging an atmosphere of collaboration and support for colleagues and the community we are surrounded by. In particular to Junrui Di and Stephen Cristiano, friends and fellow members of my cohort for all the homework and exam help and for being there when there was no else to turn to.

Thank you to Dr. Natalie Blades, a gifted teacher and an amazing mentor who introduced me as an undergraduate to the special work being done here at Johns Hopkins, as well as to many of the faculty and students.

Thank you to Dr. Jennifer Schrack and Dr. Adam Spira for being on my thesis committee and providing advice and feedback during my time here as a student.
I would like to thank Dr. Kathleen Merikangas and everyone at the National Institute of Mental Health that I had the pleasure of working with. I learned so much from collaboration on various projects that helped me broaden my understanding of public health and provided funding as an intramural pre-doctoral fellow at the NIH.

Thank you to Dr. Vadim Zipunnikov for advising, mentoring, and helping me learn so much. Your passion and excitement for statistics is infectious and has helped me get through the roadblocks and ruts when I didn’t know what to do next.

And finally, I would like to thank Dr. Ciprian Crainiceanu for allowing me to work and learn from him as an intern, and again, allowing me to work with and learn from him again as a doctoral student. You taught me many things, from the small and simple like the minutia of proper technical writing and budgeting time, to more broad statistical concepts contributing to my knowledge and understanding of the world of science; I know that your example on both ends of the spectrum will prove invaluable in the future.
# Table of Contents

Table of Contents vii

List of Tables xi

List of Figures xiii

1 Introduction 1

1.1 Predicting multiple sclerosis lesion development using longitudinal magnetic resonance imaging 1

1.2 Fragmentation and Ecological Momentary Assessment 3

1.3 Variable Domains 5

2 Dynamic prediction of multiple sclerosis lesion MRI intensities 9

2.1 Introduction 9

2.2 Data 12

2.2.1 Image acquisition and preprocessing 12

2.2.2 Data Structure 13

vii
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.3 Framing the problem</td>
<td>16</td>
</tr>
<tr>
<td>2.2.4 Demographics</td>
<td>18</td>
</tr>
<tr>
<td>2.3 Methods</td>
<td>19</td>
</tr>
<tr>
<td>2.3.1 Relative intensity normalization</td>
<td>19</td>
</tr>
<tr>
<td>2.3.2 Dynamic modeling and fitting</td>
<td>23</td>
</tr>
<tr>
<td>2.4 Results</td>
<td>27</td>
</tr>
<tr>
<td>2.4.1 Quantification and comparison of prediction accuracy</td>
<td>27</td>
</tr>
<tr>
<td>2.4.2 Prediction Model Inference</td>
<td>31</td>
</tr>
<tr>
<td>2.4.2.1 Model coefficient estimators</td>
<td>31</td>
</tr>
<tr>
<td>2.4.2.2 Visualization of estimator effects on dynamic prediction</td>
<td>35</td>
</tr>
<tr>
<td>2.5 Visualization and software</td>
<td>41</td>
</tr>
<tr>
<td>2.5.1 Results for two lesions</td>
<td>41</td>
</tr>
<tr>
<td>2.5.2 Visualization via Shiny in R</td>
<td>45</td>
</tr>
<tr>
<td>2.5.3 R code and vignette</td>
<td>48</td>
</tr>
<tr>
<td>2.6 Discussion</td>
<td>52</td>
</tr>
<tr>
<td>3 Fragmentation as a Novel Measure of Stability in Normalized Trajectories of Mood and Attention Measured by Ecological Momentary Assessment</td>
<td>58</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>58</td>
</tr>
<tr>
<td>3.2 Methods</td>
<td>61</td>
</tr>
</tbody>
</table>
3.2.1 Data ......................................................... 61
3.2.2 Procedures .............................................. 61
3.2.3 Ecological Momentary Assessment (EMA) .......... 62
3.2.4 Participant-Level Normalization ....................... 63
3.2.5 Fragmentation Measures ............................... 66
3.2.6 Missing Data ............................................ 72
3.2.7 Data Modeling .......................................... 74
3.3 Results ....................................................... 75
3.4 Discussion .................................................. 82

4 Variable Domain Functional Principal Component Analysis 89
  4.1 Introduction ............................................. 89
  4.2 Methods ................................................ 93
    4.2.1 Variable-Domain Approach ..................... 93
    4.2.2 Estimation ....................................... 95
    4.2.3 Alternative Approaches ....................... 98
  4.3 Simulation Study ....................................... 100
    4.3.1 Simulation Design ............................... 100
    4.3.2 Simulation Results .............................. 104
  4.4 Applications .......................................... 108
    4.4.1 ICU Data ......................................... 109
4.4.2 Accelerometry Data ........................................... 115
4.5 Discussion ......................................................... 122
4.6 Appendix ......................................................... 126

5 Discussion and Conclusion 133
  5.1 Predicting multiple sclerosis lesion development using longitudinal magnetic resonance imaging ........................................... 133
  5.2 Fragmentation in Ecological Momentary Assessment .............. 134
  5.3 Variable Domains ............................................... 135
  5.4 Conclusion ......................................................... 136
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Excerpt of data, showing structure for example lesions and voxels.</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>RMSE for first prediction point for models</td>
<td>28</td>
</tr>
<tr>
<td>2.3</td>
<td>RMSE for each DLM</td>
<td>29</td>
</tr>
<tr>
<td>2.4</td>
<td>Coefficient values ($\beta$) for DLM1 (historical model)</td>
<td>32</td>
</tr>
<tr>
<td>2.5</td>
<td>Coefficient values ($\gamma$) for DLM2 (demographic model)</td>
<td>33</td>
</tr>
<tr>
<td>2.6</td>
<td>Coefficient values ($\gamma, \beta$) for DLM7 (All)</td>
<td>35</td>
</tr>
<tr>
<td>3.1</td>
<td>Demographic Information Across Mood Disorder Subgroups</td>
<td>63</td>
</tr>
<tr>
<td>3.2</td>
<td>Average Bout Duration for Mood and Attention Inside the Standard Range (ISR) and Outside the Standard Range (OSR) by Mood Disorder Subgroup</td>
<td>70</td>
</tr>
<tr>
<td>3.3</td>
<td>Summary of Time Spent Inside the Standard Range (ISR) and Outside the Standard Range (OSR) by EMA Variable and Mood Disorder Subgroups</td>
<td>76</td>
</tr>
<tr>
<td>3.4</td>
<td>Significant Fragmentation Effects of Mood Disorder Subgroups for Attention and Mood Domains</td>
<td>77</td>
</tr>
</tbody>
</table>
3.5 Statistics for Mood Fragmentation Models . . . . . . . . . . . . . . . . 78
3.6 Statistics for Attention Fragmentation Models . . . . . . . . . . . . 79

4.1 Average root mean squared error (ARMSE$_Y$) from 500 simulations. 106
4.2 Average root mean squared error (ARMSE$_{PC}$) for the first 2 principal components from 500 simulations in each scenario. 108
4.3 Number of principal components needed to account for at least 99% of the variation for corresponding domain lengths $M$. 115
4.4 Number of principal components needed to account for at least 99% of the variation for corresponding domain length $M$. 121
2.1  **Lesion Voxel Normalization:**  
A) Intensity trajectory of one voxel with a high pre-incidence standard deviation (SD-solid black) and one voxel with a low SD (solid red) in a lesion from a FLAIR image, where $t = 0$ is the approximate time of incidence (solid vertical blue line). The dotted lines are located one SD (calculated using pre-incidence intensities) above and below each respective pre-incidence mean (black for high SD, red for low SD); the dashed horizontal line is the voxel-specific mean, colored accordingly. 
B) FLAIR intensity for the same voxels, normalized using the pre-incidence mean and stabilized SD of the intensities from A). The dotted lines are at $y = 1$, 1 standard deviation for both voxels. 
C) Slice of lesion at each visit showing the voxel (small black arrow) with a high standard deviation, shown in black in (A) and (B). The numbers above each slice indicate the time from incidence (in days). 
D) Slice of lesion at each time point showing the voxel (small black arrow) with a low SD, shown in red in (A) and (B).
2.2 **Nearest neighbors**: Visual representation of one voxel in a lesion. Blue cube represents the voxel specified, the six red cubes immediately surrounding the blue are the voxels used to calculate the average intensity of the “nearest” voxels in the spatial model.

2.3 **Historical Variables**: Demonstration and comparison of historical effects for DLM1 and DLM7 models. The four plots on the top row are from DLM1 (A-D) and those on the bottom row are from DLM7 (E-H). The dashed line in each plot ($-60 \leq t \leq 60$) is an example of the trajectory for one voxel used to predict future intensities in each model. The solid lines in each plot show the predictions for each model ($90 \leq t \leq 330$). The corresponding dotted lines are the 95% confidence intervals for the mean predicted values. Each color corresponds to a different “hypothetical individual/voxel” for the explanatory variables. For example, in the first plot of each row for both DLM1 and DLM7, the only explanatory variables that are changed are at $t = -60$ and $-30$. The red dashed curve has ($+1$) higher intensity at both time points, the black curve has ($-1$) lower for both, and the blue curve is in the middle. The corresponding predictions are shown in the same color. In the third plot of each row for DLM1 (C) and DLM7 (G), the only difference is at $t = 30$ where the differences are again ($+1$) and ($-1$) for the black and red curves. The remaining plots are similar and show predictions when the historical intensities are perturbed at specified time points.
2.4 **Demographic Variables:** Demonstration and comparison of demographic effects for DLM2 and DLM7 models. The four plots on the top row are from DLM2 (A-D) and the four on the bottom row are from DLM7 (E-H). The dashed line in each plot ($-60 \leq t \leq 60$) is an example of the trajectory for one voxel used to predict future intensities in each model. The solid lines in each plot show the predictions for each model ($90 \leq t \leq 330$). The corresponding dotted lines are the 95% confidence intervals for the mean predicted values. Each color corresponds to a different “hypothetical individual/voxel” for the explanatory variables. The hypothetical individuals represented by each curve have a unique value for one covariate, (e.g. age, sex) and are identical for all other covariates. Covariates examined include age, sex, EDSS score, and disease subtype (RRMS or SPMS). For example, in the first plot on each row for both DLM2 and DLM7, the only covariate that is changed is age. The three curves compare predicted trajectories (solid lines) for individuals that are 25, 35, and 45 years of age (red, blue, black curves respectively) and have all other characteristics in common. The subsequent plots are similar, but vary according to the other three covariates.
2.5 **Treatment Variables:** Treatment effects for DLM2 and DLM7 models. The two plots on the left are from DLM2 (A,B) and the two on the right are from DLM7 (C, D). The dashed line in each plot \((-60 \leq t \leq 60)\) is an example of the trajectory for one voxel used to predict future intensities in each model. Similar to Figure 2.4, we compare voxels where the individual was undergoing treatment (A, C) or taking steroids (B, D) at the time of incidence (blue is voxels that were taking treatment or steroid, red is voxels that were not).

2.6 **Prediction Example 1:** Example of lesion slice comparing predicted values from DLM7 to original interpolated values. A) and B) show original data in black/gray and predicted values in red, each curve is one voxel from the lesion slice. A) highlights a single voxel in lowest decile of RMSE (0.322) and B) highlights another voxel in the same lesion with RMSE also in the lowest decile (0.416). C) and D) show the lesion slices at each visit included in the model, historical and future. C) shows the actual normalized intensities and D) shows the predicted values. The voxels are colored according to intensity of the voxel from low (blue) to high (red). The voxels highlighted in A) and B) are indicated with arrows in both rows and labeled in the plot of the lesion at the first visit in the row.
2.7 **Prediction Example 2:** Example of lesion slice comparing predicted values from DLM7 to original interpolated values. A) and B) show original data in black/gray and predicted values in red, each curve is one voxel from the lesion slice. A) highlights a single voxel in highest decile of RMSE (3.035) and B) highlights another voxel in the same lesion with RMSE also in the lowest decile (2.308). C) and D) show the lesion slices at each visit included in the model, historical and future. C) shows the actual normalized intensities and D) shows the predicted values. The voxels are colored according to intensity of the voxel from low (blue) to high (red). The voxels highlighted in A) and B) are indicated with arrows in both rows and labeled in the plot of the lesion at the first visit in the row.

2.8 **Shiny Application:** Screenshots from Shiny application for lesion visualization. There are three tabs, the first tab on the left shows a FLAIR intensities in a lesion through each slice of the lesion, using both the original data standardized by the NAWM and the data normalized by the pre-incidence intensities. The second tab focuses on individual slices, showing the trajectories from all voxels in the slice as well as the predicted values from the model. On the bottom of the screenshot, a series of one lesion slice is shown for each visit with the original, normalized, and predicted intensities. The third panel (bottom row) shows the lesion in three dimensions, viewing one visit at a time over the course of individual follow-up.
3.1 **EMA: Data Example and Transformation:** Raw, normalized, and dichotomized transformations of exemplary Ecological Momentary Assessment (EMA) attention scores from two participants. With respect to the raw data, low levels (1 to 3) correspond to a distracted state and high levels (5 to 7) correspond to a focused state. For the binarized data, ISR (inside standard range) corresponds to normalized assessments that are within one standard deviation of the participant mean, whereas OSR (outside standard range) corresponds to normalized assessments that are greater than one standard deviation away from the mean.  

3.2 **EMA Variable Distributions:** Panel A: Distribution of raw mood scores for all participants. Panel B: Distribution of mood scores after normalization using participant-specific means and variances.  

3.3 **Fragmentation Metric Distributions:** Distribution of calculated $\lambda$s by mood disorder subgroup. The upper and lower left panels are $\lambda_0$ and $\lambda_s$ for attention respectively, and the upper and lower right panels are the densities of $\lambda_0$ and $\lambda_s$ for mood.
3.4 **Large MSSD Examples:** These two participants have high mean squared successive difference (MSSD) but different fragmentation levels ($\lambda_o$). The participant on the left is less likely to transition out of the OSR (outside the standard range) state whereas the second participant on the right tends to spend shorter amounts of time OSR and is more likely to transition out of that state. However, both participants’ raw scores are highly variable and the MSSD is approximately the same for both.

3.5 **Small MSSD Examples:** These two participants have low mean squared successive difference (MSSD) but different fragmentation levels. The first participant on the left has a high $\lambda_s$ and is more likely to transition out of the standard range than the second participant on the right, who has a low $\lambda_s$ and is more likely to stay in the standard range.

3.6 **Attention Fragmentation and MSSD:** Comparison of outside the standard range (OSR) and inside standard range (ISR) fragmentation measures (top left). Comparison of OSR fragmentation measure and mean squared successive difference (MSSD; top right). Comparison of ISR fragmentation measure and MSSD (bottom left). The correlation between the variables on the $x$- and $y$-axes is shown above each plot on the top line. The slopes and respective p-values of the fitted lines for each mood disorder subgroup are on the second line above each plot.
3.7 Mood Fragmentation and MSSD: Comparison of outside the standard range (OSR) and inside standard range (ISR) fragmentation measures (top left). Comparison of OSR fragmentation measure and mean squared successive difference (MSSD; top right). Comparison of ISR fragmentation measure and MSSD (bottom left). The correlation between the variables on the $x$– and $y$-axis is shown above each plot on the top line. The slopes and respective p-values of the fitted lines for each mood disorder subgroup are on the second line above each plot. 

4.1 SOFA scores for individuals with acute lung injury (ALI) in the ICU: A: Patients are divided into two groups, those that died (bottom) and those that survived and were discharged (top). Each row in the plot depicts the SOFA scores measured once daily for each subject and subjects are ordered by the length of time in the hospital, darker red indicates higher SOFA scores. B: SOFA score trajectories from five subjects are highlighted in black. Scores for all other subjects are shown in gray. C: Estimated empirical distribution of length of stay on log scale for all subjects; patients who survived and were discharged are represented by the distribution in red and patients who died are represented by the distribution in green. 

xx
4.2 **Mean and Covariance Surface Depictions:** Comparison of the mean and covariance surfaces in the variable domain (VD) case in the right column and the same domain (SD) case in the left column. The two figures in the top row are representations of conditional mean surfaces. Likewise, the two figures in the bottom row are the corresponding representations of the covariance function space, with the VD case on the right.

4.3 **Eigenfunction Transformations:** 

**A:** First eigenfunction \((k = 1)\) linearly stretched (Method A) over domain lengths \(m = 10, 30,\) and 60. 

**B:** Second eigenfunction \((k = 2)\) linearly stretched (Method A). 

**C:** First eigenfunction \((k = 1)\) using weighted combination of two eigenfunctions (non-linear, Method B). In C, the eigenfunctions transform from the second eigenfunction in B at \(m = 10\) to the first eigenfunction in A at \(m = 60\).

4.4 **Simulation Setup:** 

**A:** Trajectories for 500 simulated subjects. Four individual subjects are highlighted in various colors. Mean function is shown with dashed black line. 

**B:** Overlay of the two distributions of \(M\) (D1 and D2) and the weighting function \(W(m)\). D1, shown in red, is a Uniform(2, 80). D2, shown in gray, is Geometric(0.06).
4.5 **Variable Domain PCs from ICU data:** Left: Heatmaps of first and second principal component functions from vd-FPCA method across all domain lengths of the ICAP data from 2 to 170 days, ordered by domain length. Right: Selected examples of first and second principal components for \( m = 10, 60, 110, \) and 160 days. This shows the change in principal components as \( M \) increases is more complicated than linearly scaling to the domain length. In parentheses above each plot is the percentage of total variance accounted for by the component.

4.6 **Example Participants by PC scores:** Daily SOFA levels for subjects with hospital stays of length \( m = 10, 60, 110, \) and 160 days in each column. Black solid line shows SOFA for subject with a high (positive) PC score, red solid line shows subject with corresponding low (negative) score and dashed blue line shows the mean SOFA levels for each domain length \( M \). The top row corresponds to PC 1, and the bottom row to PC 2. The plot for PC 2 for \( m = 10 \) days was excluded because the first PC accounts for more than 99% of the variation between subjects.
4.7 **Scaled and Weighted Principal Component Functions:** Scaled is on the left and weighted is on the right. The PCs account for at least 99% of the variance for each approach. The functions are quite different across the various approaches and the scaled and weighted approaches appear to be accounting for noise with the later components (PCs 5 and 6).

4.8 **Variable Domain PCs from Accelerometry Data:** Left: Heatmaps of first and second principal component functions from vd-FPCA method across all domain lengths of the accelerometry data from 0.5 to 2.5, ordered by domain length. Right: Selected examples of first and second principal components for \( m = 0.7, 1.2, 1.7, \) and 2.2 seconds. In parentheses above each plot is the percentage of total variance accounted for by the component. This shows that change in principal components as \( M \) increases is more complicated than linearly scaling to the domain length.

4.9 **Example Participants by PC scores:** Acceleration for subjects standing from a chair. Subjects with high (black solid line) and low (red solid line) corresponding PC scores are compared at domain lengths \( m = 0.7, 1.2, 1.7, \) and 2.2 seconds. The mean for each domain length is shown in blue (dashed line). Plots on the top row (A-D) correspond to PC 1 and plots on the bottom row (E-H) correspond to PC 2.
4.10 **Scaled and Weighted Principal Component Functions**: Principal component functions using the scaled (left) and weighted (right) methods. The PCs account for at least 95% of the variance for each approach. The functions vary across approaches, the scaled approach needs more components to account for the variation and later PCs appear to model noise artifacts.

4.11 **Non-linear Transformations**: The black dashed line shows the original function $m(x, j)$ from Equation (8) where $j = 6$. The red line and corresponding points show the same function with the transformed domain $T(x) = x^\alpha$ for $\alpha = (0.1, 0.5, 2, 10)$.

4.12 **Smoothing with Non-linear Transformations**: Smooth estimates of a function with and without transformation of the domain. Raw data is shown in dark gray and the smoothed data on the original domain is shown in red. The green and blue curves show the smooth estimates after transformation of the domain, $x^\alpha$ for $\alpha = 0.01$ and 20 respectively.
Chapter 1

Introduction

This thesis is laid out in the following chapters: 1) Dynamic prediction of MRI intensities in a longitudinal study of individuals with multiple sclerosis (MS), 2) Fragmentation in normalized trajectories of mood and attention measured by ecological momentary assessments (EMA), and 3) Variable-domain functional principal component analysis.

1.1 Predicting multiple sclerosis lesion development using longitudinal magnetic resonance imaging

Throughout the course of disease progression in multiple sclerosis (MS), lesions develop in the brain and spinal cord, damaging nerve tissue and harmfully impacting the body. The development and atrophy of these lesions has been shown to be related to disease development (Dwyer et al., 2018). After lesion incidence, some lesions, or regions of lesions, recover similar to their appearance before the lesion formed. Other regions or entire lesions do not and after an initial decrease in
inflammation never recover and permanently damage the brain. In chapter 2, we introduce a class of methods to predict how the lesion develops after it is formed and if or when it will be "repaired" and stabilized. We do so dynamically, using the lesion history up to a given time, predicting what the lesion will look like at any particular time in the future.

As part of a natural history study at the National Institute of Neurological Disorders and Stroke (NINDS), 60 individuals with MS received brain scans using structural magnetic resonance imaging approximately monthly for up to 5.5 years. Since the time between scans is short (median of 28 days) and the number of scans is large (22.5 per person on average), lesions can be tracked at a fairly high resolution. However, the time between each scan is highly variable, they can be as few as two weeks apart or have up to a year between scans. This irregularity between scans is compounded due to the fact we do not know precisely when the lesions form. Lesion incidence could be one day before the scan, or more than 30 days which could drastically effect the accuracy of the model and predictions.

To provide a complete and integrated solution, three important methodological problems were addressed: (1) relative normalization of the arbitrary intensity units in sMRI to facilitate the comparison of the current lesion condition to pre-incidence levels; (2) the spatial localization and mutual correlations of voxels in the lesion; and (3) the parameterization of prediction models, which depends on the prediction timing (when) and horizon (how far into the future) prediction is conducted as well as on the amount of available historical information.
The unique structure and size (751 lesions from 48 participants) of the data from this study also provides an opportunity to develop visual tools that facilitate the interactive exploration of the data and corresponding model results. We created an interactive application in Shiny that enables the visual examination of lesions from various perspectives, including: 1) a comprehensive view of the entire lesion and its development over time, 2) a narrow point of view, focusing on different slices or regions of the lesion across time, including comparisons of model results, and 3) three-dimensional representations of the lesions at any given point of time. The methods we showcase here are important steps toward understanding lesion behaviour and predicting if or when brain lesions will return to their baseline states.

1.2 Fragmentation and Ecological Momentary Assessment

Ecological momentary assessment (EMA), also referred to as daily diaries or the experience sampling method (ESM), is a versatile research methodology that seeks to increase ecological validity and minimize recall bias by having participants answer brief surveys repeatedly, often multiple times per day. Through mobile devices like smartphones, assessments can be filled out throughout the day practically anywhere and can be regularly scheduled, triggered by certain events, or taken randomly throughout the day depending on the study design. They have the potential to explore dynamic relationships between an individual’s experiences, environment, and responses, including fluctuations within time periods as short as
a day (Ebner-Priemer et al., 2009). All of the unique features and flexibility that come with EMA introduce interesting complications and opportunities regarding the treatment and analysis of the results as well.

Historically, EMA has provided novel insights into diverse psychological processes. One such area which we explore is the stability of emotional states in subgroups of individuals with mood disorders, specifically, Bipolar I (BPI), Bipolar II (BPII), and major depressive disorder (MDD), compared to a control group with no lifetime history of these disorders. Participants in the mobile technology component of a large community-based family study through the National Institute of Mental Health were asked to complete four assessments each day at equally spaced intervals tailored to each individual’s schedule for two weeks. The assessments included a variety of questions concerning mood and cognitive states, daily event occurrences, and diverse additional variables assessing behavior or experiences. In this investigation, we focus on one self-reported mood state and one self-reported cognitive state question.

Analyses of these data typically have focused on modeling participant-specific averages, variability, and stability. Fragmentation analysis, a method of quantifying stability, was originally developed to characterize patterns of sedentary time accumulation in objectively measured physical activity and has been explored in depth for physical activity measures (Di et al., 2017). However, it has not been used in EMA applications, which is similar in structure to physical activity analysis, but its unique structure requires some modifications.
In chapter 3, we propose a novel statistical framework to determine participant
stability by quantifying fragmentation of standardized trajectories using the fol-
lowing 2-step approach: (1) participant-level EMA scores are normalized, and (2)
normalized scores are dichotomized into 2 states, inside and outside a range of
1 standard deviation. Within-participant fragmentation measures are then calcu-
lated from the dichotomized scores and modeled with various covariates. We use
this method to study patterns of emotional states and showed that the proposed
fragmentation measures differentiate mood disorder subtypes, including Bipolar I
(BPI), Bipolar II, and major depressive disorder (MDD) compared with unaffected
controls. Fragmentation measures were regressed on the mood disorder subtype,
adjusting for age, sex, body mass index, and mean squared successive difference.

1.3 Variable Domains

Functional principal component analysis is a common approach for identifying the
main directions of variation in functional data. However, traditional approaches
assume for the most part that the functional domain for each individual is the same.
This can be problematic for many reasons when this is not the case. Our motivating
example comes from a study on Improving Care of Acute Lung Injury Patients
(ICAP) where daily assessments of overall health status were performed for 520
participants (Needham et al., 2006). The Sequential Organ Failure Assessment
(SOFA) scores were collected during their stay in the hospital Intensive Care Unit
(ICU) up until death or discharge from the hospital. For some individuals, this
meant only a few days, in fact some were in the ICU for as few as one day. Others remained in the hospital for multiple months, some more than 150 days.

If we wanted to find the main directions of variation, or perform principal component analysis using this study, an alternative to the brute force approach would be necessary. One straightforward alternative in cases where the number of observations for each participant’s trajectory is not the same that is not uncommon is to transform the data to the same domain, interpolate the data onto a common grid, and conduct PCA on the transformed data. This would mean we are treating the relative time an individual has been in the ICU relative to their length of stay. Specifically, if we compare the SOFA score trajectory in an individual that was in the ICU four days, to an individual who stayed for 100 days, we would be comparing day two for the first individual and day 50 for the second as if they were the same for each individual, which makes little sense scientifically. Additionally, in-patient stays in the hospital are not the only case where functional domains can vary, extremely or otherwise. Variable-domain functional data are becoming increasingly common in health studies. EEG records of brain activity during sleep will vary by the duration of each individuals sleep. Acceleration profiles for different movements can vary between individuals and even within individuals. Continuous observation of study participants during follow up can vary substantially as well.

In chapter 4, we propose a functional principal component technique for variable-domain data, allowing components to vary according to domain length. To do so, we fit a trivariate smoother using penalized thin plate splines to estimate
the covariance as a function of the domain length. Then we condition on the 
domain length and calculate the principal components eigen-decomposition of 
the estimated covariance matrix. To assess this approach, we apply it in two real 
data settings with variable functional domains as well as multiple simulated data 
scenarios. This approach is straightforward with standard software ready to use 
and is quite flexible, allowing principal components to vary according to the length 
of the functional domains.
References


Chapter 2

Dynamic prediction of multiple sclerosis lesion MRI intensities

2.1 Introduction

Multiple sclerosis (MS) is a disease of the central nervous system characterized by demyelinating lesions in the brain and spinal cord. The number and total volume of lesions have been reported to be associated with adverse health effects and quality of life measures (Hohol et al., 1997; Khoury et al., 1994; Sperling et al., 2001; Rudick et al., 2007). Lesions typically develop from active regions of tissue inflammation to stable, demyelinated areas of axonal injury over the course of several months (Lassmann, Bruck, and Lucchinetti, 2007; Lassmann, 2013). Lesion atrophy has also been linked to changes in the expanded disability status score (EDSS), a common indicator of the individual’s functional disability (Dwyer et al., 2018). Remyelination, a competing process, occurs at varying efficiency levels for each individual and is documented in both relapsing-remitting and secondary progressive MS cases (Patrikios et al., 2006; Bramow et al., 2010). The
dynamics of lesion development, including the destructive (lesion expansion, tissue inflammation, and demyelination) and restorative (tissue inflammation reduction and remyelination) processes could be related to an individual’s inherent ability to respond to MS disease dynamics and may be modifiable by treatment.

White matter lesions appear hyper-intense on fluid attenuated inversion recovery (FLAIR), T2-weighted (T2w), and proton density (PD) images and hypo-intense in T1-weighted (T1w) images. Once a lesion occurs, its multi-sequence sMRI voxel intensities change continuously and this dynamic process contains two distinct phases. In the first phase, the lesion expands to its maximum size and is associated with inflammation. In the second phase, the inflammation typically decreases and the image hyper-intensities of individual voxels in the lesion either remain the same or slowly decrease. The intensity of some voxel intensities return to their baseline distribution (before lesion formation) and are thus indistinguishable from a healthy voxel, at least on multi-sequence sMRI. Other voxels do not return to their baseline distribution, but most tend to stabilize around a particular level of intensity.

Due to the clinical significance of lesion development in individuals with MS, we investigate how dynamic lesion changes can be predicted from their history as measured by multiple sMRI sequences and other covariates. Previous solutions to this problem have used retrospective (static) models of longitudinal sMRI. For example, Meier and Guttmann, 2003; Meier and Guttmann, 2006 and Meier, Weiner, and Guttmann, 2007 showed that: (1) the largest change between voxel intensities before and after lesion occurrence is observed closer to the center of the lesion; (2)
lower initial intensity of a lesion voxel was predictive of repair; and (3) most lesion activity did not last beyond 10 weeks after the initial lesion formation. Ghassemi et al., 2015 investigated the association between changes in sMRI normalized T1-weighted intensity in new lesions and disease duration and treatment status.

In a recent paper, Sweeney et al., 2016 showed that changes in the intensity of lesion voxels measured by multi-sequence sMRI are associated with the distance to the boundary of the lesion, age, and treatment. Dworkin et al., 2016 developed methods to predict the appearance of lesions on multi-sequence sMRI one year post-incidence using clinical, demographic, and imaging information from a single visit at incidence. Building on this research, we focus on using sMRI scans up to a predetermined time after lesion incidence to predict: (1) the intensity for any voxel in the lesion at any time into the future; and (2) the time it takes for a lesion voxel intensity to recover to pre-incidence levels. We examine a class of dynamic regression models and provide quantitative and visualization tools for characterizing the performance of our models. These tools are useful for characterizing if and how lesions respond to treatment, at least as assessed by multi-sequence sMRI. These results could be further used to inform clinical trial design and sample size calculations as well as provide information regarding which treatment regimens for individuals may be most effective throughout the course of disease development.
2.2 Data

To illustrate some aspects of the complexity of predicting MS lesion behavior, we include an in-depth look at the data setup and structure. We first introduce the study image acquisition and preprocessing details and then review the data structure and visualization.

2.2.1 Image acquisition and preprocessing

Complete details concerning image acquisition and pre-processing have been published previously in Sweeney et al., 2016, but are summarized here, as well. A total of 60 study participants diagnosed with multiple sclerosis (MS) were scanned approximately monthly between 2000 and 2008. Study participants were scanned as part of a natural history study at the National Institute of Neurological Disorders and Stroke (NINDS) for a period of up to 5.5 years (mean = 2.2 years, SD = 1.2). Whole brain $T_2$, FLAIR, and PD (2D) as well as $T_1$ (3D) images were obtained from a 1.5 tesla MRI scanner (Signa Excite HDxt; GE Healthcare, Milwaukee, Wisconsin). Preprocessing was done using Medical Image Processing Analysis and Visualization (http://mipav.cit.nih.gov) and the Java Image Science Toolkit (Lucas et al., 2010). Images were interpolated to a voxel size of 1mm$^3$ and rigidly co-registered across time and sequence to a template space (Fonov et al., 2011). A two step procedure was applied to co-register the $T_1$ images across visits: first, subject-specific templates were generated by averaging the $T_1$ images after rigid registration to the MNI template. Second, all $T_1$ images were registered to the subject-specific $T_1$
templates. The other MRI sequences were then aligned to the T\textsubscript{1} images within each study visit. Co-registered (within-subject) images were skull-stripped (Carass et al., 2007) and an automatic segmentation algorithm (Shiee et al., 2010) based on T\textsubscript{1} and FLAIR images was used to produce a mask of normal appearing white matter (NAWM). Intensity normalization was then conducted using z-scoring based on the mean and variance of the intensities in the NAWM (Shinohara et al., 2014). After preprocessing, images were quality controlled manually by a researcher with over five years of experience in sMRI analysis and studies with motion or other artifacts were removed. Lesions in the brain were automatically identified and segmented using a combination of OASIS and SuBLIME methods (Sweeney et al., 2013a; Sweeney et al., 2013b; Sweeney et al., 2016).

### 2.2.2 Data Structure

For the purpose of this paper we will use only the extracted longitudinal lesion data obtained via the preprocessing pipeline described in Section 2.2.1. The dataset contains 48 patients with 751 lesions that contained 187,000 voxels with an average of 21 observations per trajectory. Fundamentally, each lesion is identified as a set of voxel locations, which are then tracked over time. For each voxel location we have the sMRI sequence intensity (T\textsubscript{2}, FLAIR, PD (2D) and T\textsubscript{1}).

Table 2.1 provides the structure of the data (we only show FLAIR intensities to reduce the visual complexity of the table). The first column is the patient ID (PID) and the second column indicates lesion ID (LID), as the same patient may have multiple lesions through the course of the study. The third column indicates
Table 2.1: Excerpt of data, showing structure for example lesions and voxels.

<table>
<thead>
<tr>
<th>PID</th>
<th>LID</th>
<th>VID</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Visit</th>
<th>Days FSS</th>
<th>Days FI</th>
<th>FLAIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>1</td>
<td>0</td>
<td>-168</td>
<td>-0.81</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>2</td>
<td>28</td>
<td>-140</td>
<td>0.11</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>3</td>
<td>56</td>
<td>-112</td>
<td>-0.22</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>4</td>
<td>77</td>
<td>-91</td>
<td>0.47</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>5</td>
<td>112</td>
<td>-56</td>
<td>-0.22</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>6</td>
<td>126</td>
<td>-42</td>
<td>0.99</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>7</td>
<td>168</td>
<td>0</td>
<td>2.57</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>8</td>
<td>182</td>
<td>14</td>
<td>1.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>9</td>
<td>196</td>
<td>28</td>
<td>1.24</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>10</td>
<td>210</td>
<td>42</td>
<td>1.27</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>11</td>
<td>238</td>
<td>70</td>
<td>1.94</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>17</td>
<td>434</td>
<td>266</td>
<td>0.21</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>18</td>
<td>483</td>
<td>315</td>
<td>1.84</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>123</td>
<td>120</td>
<td>99</td>
<td>1</td>
<td>0</td>
<td>-168</td>
<td>-0.07</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>123</td>
<td>120</td>
<td>99</td>
<td>2</td>
<td>28</td>
<td>-140</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>73</td>
<td>73</td>
<td>103</td>
<td>113</td>
<td>7</td>
<td>168</td>
<td>-42</td>
<td>1.86</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>73</td>
<td>73</td>
<td>103</td>
<td>113</td>
<td>8</td>
<td>182</td>
<td>-28</td>
<td>1.04</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>73</td>
<td>73</td>
<td>103</td>
<td>113</td>
<td>9</td>
<td>196</td>
<td>-14</td>
<td>1.04</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>73</td>
<td>73</td>
<td>103</td>
<td>113</td>
<td>10</td>
<td>210</td>
<td>0</td>
<td>5.10</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>73</td>
<td>73</td>
<td>103</td>
<td>113</td>
<td>11</td>
<td>238</td>
<td>28</td>
<td>1.58</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>73</td>
<td>73</td>
<td>103</td>
<td>113</td>
<td>12</td>
<td>266</td>
<td>56</td>
<td>1.56</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>3466</td>
<td>73</td>
<td>91</td>
<td>108</td>
<td>1</td>
<td>0</td>
<td>-63</td>
<td>2.30</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>3466</td>
<td>73</td>
<td>91</td>
<td>108</td>
<td>2</td>
<td>28</td>
<td>-35</td>
<td>3.55</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>3466</td>
<td>73</td>
<td>91</td>
<td>108</td>
<td>3</td>
<td>63</td>
<td>0</td>
<td>4.51</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>3466</td>
<td>73</td>
<td>91</td>
<td>108</td>
<td>4</td>
<td>84</td>
<td>21</td>
<td>4.37</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>3466</td>
<td>73</td>
<td>91</td>
<td>108</td>
<td>5</td>
<td>105</td>
<td>42</td>
<td>4.35</td>
</tr>
</tbody>
</table>

Note: PID = Participant ID; LID = Lesion ID; VID = Voxel ID; Visit FI = Visit number where 0 is the first visit the lesion is evident; Days FSS = Days from study start where 0 is the first day of the first visit; Days FI = Days from lesion incidence where 0 is the day of the visit the lesion is first evident; FLAIR = sMRI FLAIR intensity after preprocessing.

The voxel ID number is unique for all voxels, meaning each voxel from any particular lesion will have a different ID than all voxels from all other lesions.
the voxel ID (VID) and the fourth through sixth columns show the X, Y, and Z coordinates of the voxel in the patient, brain-specific reference frame. These columns are useful to understand the relative location of the lesion voxels as well as their position relative to the center of the lesion. The seventh and eight columns are the visit indicator and time expressed in days since entering the study. The ninth column (Days FI) provides information about the number of days before the visit when the lesion was identified (indicated by a negative number) or after the lesion is detected (indicated by a positive number). For example, for PID = 1, the lesion with LID = 1 was detected at Visit 7 (note that the corresponding Days FI variable is 0), when the normalized FLAIR intensity jumped from values between −0.81 (at Visit 1, 168 days before the lesion was detected) and 0.99 (at Visit 6, 42 days before the lesion was detected) to 2.57 (see the column labeled FLAIR and the corresponding row at Visit 7). After the initial visit, the normalized FLAIR intensities of this voxel decrease again to 1.10 at Visit 8, 14 days after lesion detection, and 1.24 at Visit 9, 28 days after lesion detection. Recall that normalized FLAIR intensity images are expressed in standard deviations of the NAWM intensities. Larger, positive values indicate image intensities above the average NAWM intensity, while values above 2 tend to be visually identifiable as hyper-intense observations, as the density of NAWM voxels is close to a normal density.

The second lesion for the first patient (LID = 2 for PID = 1) was detected at Visit 10, 210 days into the study (note the corresponding Days FSS = 210 and Days FI = 0). In contrast, the first lesion was first detected at Visit 7, 168 days
into the study. In this data, the first 72 voxels (Voxel ID 1 – 72) were part of the first lesion (LID = 1). Thus, the first voxel for PID = 1 for LID = 2 has VID = 73. For this voxel the jump in the normalized FLAIR intensity is larger. Indeed, the FLAIR intensities at Visits 7, 8, and 9 are 1.86, 1.04, and 1.04 while at visit 10 the normalized intensity jumps to 5.10 standard deviations, a value of intensity that would be easy to recognize by visual inspection as being strongly hyper-intense. The last part of Table 2.1 shows information about the first voxel (VID = 3466) from the 25th lesion (LID = 25) in the study, which was observed in the third patient (PID = 3). For this individual, the lesion labeled LID = 25 occurred on the third visit only 63 days after the start of the study.

2.2.3 Framing the problem

Given the data structure introduced in Section 2.2.2 and illustrated in Table 2.1, we are interested in predicting future voxel-specific image intensity trajectories using historical information at the voxel level. We refer to this problem as dynamic prediction (Ivanescu, Crainiceanu, and Checkley, 2017). For example, for PID = 1, LID = 1, VID = 1 one might be interested in predicting the FLAIR intensities at all time points after Visit 9, which occurred 28 days after lesion incidence. To achieve this, we use all information available before Visit 9 as well as the association between the past and future trajectories of all voxels at all lesions, where the present is defined as day 28 after lesion incidence. However, because there is nothing special about 28 days after lesion incidence, we are interested in producing such predictive models at every time point after lesion incidence.
Dynamic (prospective) modeling and prediction is fundamentally different from static (or retrospective) modeling (Meier and Guttmann, 2003; Meier and Guttmann, 2006; Sweeney et al., 2016), which is typically applied after all data are collected. Indeed, dynamic modeling allows continuous updating of predictions as new imaging data become available and can provide quantification and ranking of the effects of important predictors as a function of time. For example, one can quantify the importance of age relative to the last voxel image intensity to predict the lesion voxel intensity at any point into the future. This makes dynamic prediction highly relevant in the clinical context, as predictions are tailored to the patient-specific data, while using available information from other lesions and patients. The closest approach to what we propose here is the one by (Dworkin et al., 2016), who used longitudinal lesion profiles at the time of lesion incidence to predict future lesion recovery a year after incidence. However, to the best of our knowledge, none of these approaches are designed to dynamically predict the future voxel intensity trajectories on a dense grid of time points. Therefore, we propose to address this gap in the literature by introducing dynamic prediction methods for future voxel-level trajectories at any time after lesion incidence conditional on nearby voxel intensities, history, treatment, and demographic variables.

The most closely related methodological approaches have been published recently Ivanescu, Crainiceanu, and Checkley, 2017; Leroux et al., 2018 who proposed explicit and implicit dynamic prediction models with application to child growth data. While related, our methods are different in several different ways. First, methods need to account for the much larger size of the data as current methods
have been shown to work well with up to 200-300 individual trajectories. However, their performance has not been tested on large data sets with hundreds of thousands of trajectories, such as our sMRI example. Second, there is considerable within- and between-patient variability in the timing between subsequent visits. As an example, Table 2.1 shows that patients 1 and 3 both had their second visit 28 days after the first, but this visit is the only one where the timing of the visits align. Intervals between visits are typically about 30 days (median time is 28 days, mean is 35.8 days) and can range from around 2 weeks to about 2 months, with some intervals being as long as a year or more. Furthermore, lesion data from the same patient (such as lesions 1 and 2 in Table 2.1) do not start on the same visit; this happens because time is reset to 0 every time a new lesion is detected, which induces within-patient variability of visit times.

2.2.4 Demographics

To be included in the analysis, study participants needed at least one new lesion during the observation period. For each lesion to be included, at least 2 visits pre-incidence and 2 visits post-incidence were required. A total of 48 study participants and 751 lesions met the inclusion criteria. Study participants ranged in age from 18 to 60 years (mean age 37.6 and standard deviation 9.6 years.) The median total follow-up time for participants was 595 days (mean = 716 days) which covered approximately 21 visits on average. The average expanded disability status scale (EDSS) was 2.6. Out of the 48 study participants, 30 (63.3%) were females, and 3 (6.1%) were on steroids at the start of the study. Of the 48 subjects, at baseline
42 (87.5%) had relapsing-remitting MS (RRMS), and 6 (12.5%) had secondary progressive MS (SPMS).

2.3 Methods

2.3.1 Relative intensity normalization

Recall that the data shown in Table 2.1 is already normalized with units representing z-scores relative to the NAWM distribution. However, this normalization does not help quantify how far the voxel-specific sMRI intensities are from the distribution of its baseline intensities. Indeed, before lesion formation, the sMRI voxel intensities in the area where the lesion will appear are not spatially homogeneous. Indeed, some voxels appear brighter (e.g., higher average FLAIR intensities) and some appear darker (e.g., lower average FLAIR intensities). Thus, voxel recovery, at least in terms of sMRI, will need to be defined as a voxel-by-voxel basis.

For this reason, we consider the following re-normalization approach \( \tilde{y}_{ikv}(t_j) = \{y_{ikv}(t_j) - \hat{\mu}_{ikv}\}/s_{ikv} \), where \( y_{ikv}(t_j) \) is the z-score normalized intensity relative to NAWM of the voxel \( v \) in lesion \( k \) of study participant \( i \) at time \( t_j \) associated with visit \( j \). Here \( \hat{\mu}_{ikv} \) and \( s_{ikv} \) are the mean and standard deviation of \( y_{ikv}(t_j) \) for \( t_j < 0 \) (all times before lesion incidence), respectively. The re-normalized values, \( \tilde{y}_{ikv}(t_j) \), have the desirable property that they measure departures from the baseline voxel-specific mean intensity as a multiple of the baseline standard deviation of that specific voxel. However, for some voxels, the baseline standard deviation can be very small, especially when there are few baseline observations. Therefore, such
a re-normalization can substantially inflate the intensities $\tilde{y}_{ikv}(t_j)$ as measures of distance from the baseline distribution. To address this problem we define the following stabilized standard deviation

$$
\tilde{s}_{ikv} = \sqrt{\frac{s_{ik}^2 + s_{ikv}^2}{2}},
$$

where $s_{ik}^2 = \sum_{v=1}^{V_k} s_{ikv}^2 / V_k$, $V_k$ is the number of voxels in lesion $k$, and $s_{ikv}^2$ is the average of all baseline voxel-specific variances, $s_{ikv}^2$. Using $\tilde{s}_{ikv}$, we define the new baseline normalized intensities as

$$
\tilde{y}_{ikv}(t_j) = \frac{y_{ikv}(t_j) - \hat{\mu}_{ikv}}{\tilde{s}_{ikv}}.
$$

(2.1)

Using the stabilized standard deviation for normalization protects against small variances at the voxel level. Indeed, the average of voxel-specific variances of baseline intensities, $s_{ikv}^2$ will tend to dominate the voxel-specific variance, $s_{ikv}^2$, when $s_{ikv}^2$ is small. However, when $s_{ikv}^2$ is moderate or large, its contribution to the variance $\tilde{s}_{ikv}^2$ becomes substantial. Thus, our approach is a compromise that protects against extreme cases (small variances) and accounts for the baseline differences in variation at the voxel level.

Figures 2.1-A and B show the normalization of two example voxels, one voxel (shown in black) with a relatively large pre-incidence standard deviation (SD) and another one (shown in red) with a relatively small SD. The x-axis of both plots shows the time $t$ in days where $t = 0$ indicates the first time (visit) when the lesion was observed and $-100$ indicates 100 days before the first time the lesion
Figure (2.1) Lesion Voxel Normalization: A) Intensity trajectory of one voxel with a high pre-incidence standard deviation (SD—solid black) and one voxel with a low SD (solid red) in a lesion from a FLAIR image, where $t = 0$ is the approximate time of incidence (solid vertical blue line). The dotted lines are located one SD (calculated using pre-incidence intensities) above and below each respective pre-incidence mean (black for high SD, red for low SD); the dashed horizontal line is the voxel-specific mean, colored accordingly. B) FLAIR intensity for the same voxels, normalized using the pre-incidence mean and stabilized SD of the intensities from A). The dotted lines are at $y = 1$, 1 standard deviation for both voxels. C) Slice of lesion at each visit showing the voxel (small black arrow) with a high standard deviation, shown in black in (A) and (B). The numbers above each slice indicate the time from incidence (in days). D) Slice of lesion at each time point showing the voxel (small black arrow) with a low SD, shown in red in (A) and (B).

is detected. The $y$-axis in Figure 2.1-A shows the original z-score intensities from the FLAIR images obtained by subtracting the mean and dividing by the standard deviation of the normally appearing white matter (NAWM) intensities. Figure 2.1-B shows FLAIR intensities that have been normalized by the pre-incidence voxel...
mean and stabilized variance, as described in equation (2.1). These two figures show somewhat typical trajectories, which before lesion incidence \( t = 0 \) vary around their pre-incidence means (red and black dashed horizontal lines in Panel A). At \( t = 0 \) they both spike and then gradually decrease toward levels that are near, but just above, their pre-incidence levels. On the original scale (Panel A in Figure 2.1), 200 days after lesion detection the FLAIR intensities of the two voxels are comparable. However, when using the stabilized standard deviation, the intensities of the voxel with higher pre-incidence variance get closer to their baseline distribution. These are important distinctions with direct modeling and interpretation consequences. One of the original objectives was to determine whether and when each voxel will return to its pre-incidence intensity range. Using the voxel-specific pre-incidence mean and standard deviation to normalize each voxel provides an intuitive quantification of how far the FLAIR intensities are from their pre-incidence values, which has the same interpretation across voxels, lesions, and subjects. Dynamically predicting the voxel-specific trajectories within each lesion is an important first step to accomplishing this objective.

Figures 2.1-C and D display the FLAIR intensities of lesion slices that contain the high and low baseline variance voxels shown in panels A and B respectively (each voxel indicated by a small black arrow) for each visit up to 350 days after lesion incidence. The labels at the top of each slice indicate time from incidence (in days) of each visit. Each voxel in the two slices is colored according to the FLAIR intensity on the original scale, blue for low values \( \sim 0 \) and red for high values, in this case greater than 3. At time zero (when the lesion is detected) there is an
obvious change, both slices changing color from blue to red. Within 350 days, most voxels in the lesion in Figure 2.1-D appear to return closer to their baseline blue. In contrast, Figure 2.1-C displays a lesion that does not seem to return to the baseline color (intensity), indicating that the lesions did not recover completely. In these two examples, it is apparent that voxels closer to the edge of the lesion tend to return more quickly to their baseline levels (and color), whereas the voxels in center of the lesions seem to change more at lesion incidence and return slower to their baseline intensities. Thus, panels C and D in Figure 2.1 suggest that the position of voxels relative to the edge of the lesion can provide substantial information about their dynamic behavior.

2.3.2 Dynamic modeling and fitting

We introduce dynamic linear models (DLM), a class of richly parameterized linear models, and compare their performance as a function of historical, demographic, and spatial information (Ivanescu, Crainiceanu, and Checkley, 2017). Specifically, we assume that historical information for a lesion up to a predetermined time $t^* > 0$ after lesion incidence (e.g., $t^* = 30$ or 60 days) is available. We dynamically predict “future” voxel trajectories, and define the linear model for the normalized FLAIR intensity $\tilde{Y}_{ikv}(t)$ from subject $i$, lesion $k$, and voxel $v$ at time $t = t^* + t_s$ using historical information from time $t^* - t_h \leq t \leq t^*$ as

$$\tilde{Y}_{ikv}(t^* + t_s) = W_{ikv} \gamma_{s,t^*,h} + \sum_{t=t^*-t_h}^{t^*} \tilde{Y}_{ikv}(t) \beta_{s,t^*,h}(t) + \epsilon_{ikv}^s. \quad (2.2)$$
Here, $W_{ikv}$ is the vector of covariates for subject $i$, lesion $k$, and voxel $v$ in addition to history of the voxel FLAIR intensity. We use bold font to indicate that both $W_{ikv}$ and $\gamma_{s,t^*,h}$ are vectors. A different notation is used for the historical information, $\tilde{Y}_{ikv}(t)$, which depends both on when the prediction is conducted, $t^*$, and on all the time points between $t^* - t_h$ and $t^*$. The time $t^* - t_h$ is the earliest visit where the FLAIR intensity from voxel $v$ is used in the model. After defining $t^*$ and $t^* - t_h$, we fit a new model for each future time $t^* + t_s$ with parameter vectors $\gamma_{s,t^*,h}$ and $\hat{\beta}_{s,t^*,h}(t)$ that are uniquely defined for each current time $t^*$, amount of history $h$ used ($t^* - t_h$ to $t^*$), and future prediction time $t_s$. The voxel intensities, $\tilde{Y}_{ikv}(t)$ from visits where $t \in [t^* - t_h, t^*]$, are used to predict future voxel intensities, $\tilde{Y}_{ikv}(t^* + t_s)$, under the assumption that $e_{sikv} \sim N(0, \sigma_s^2)$ are independent.

Although time is continuous, we treat it as discrete on a pre-defined set of grid points. Visits took place approximately monthly (median time between visits was 28 days). However, the actual time interval between visits varies substantially within- and between-patients. Thus, we interpolate the FLAIR intensities at a grid of points of 30 day intervals, $t = (..., -60, -30, 0, 30, 60, ...)$, where $t = 0$ indicates the scan where the lesion is identified. For the visits from $t^* - t_h$ to $t^*$, we use linear interpolation of scans observed only at or before $t^*$ to approximate the FLAIR intensities used as covariates in the model. For visits where $t > t^*$, we use a “flat interpolation” to estimate the intensities at 30 day intervals (e.g., $t = 90, 120, 150, ...$). More specifically, all scans within a 30 day range, from 15 days before $t$ to 15 days after $t$, are binned and averaged. For example, if $t = 150$, we average the lesion intensities for all scans recorded at time $t$, where $135 \leq t \leq 165$. 

24
Given the uncertainty about the exact time of the lesion incidence (i.e., the lesion could have formed 29 days before it was actually detected \(t = -29\) or one day before \(t = -1\)) and the fact that in clinical practice scans are rarely conducted more often, it seems reasonable to conduct both modeling and prediction on a monthly grid of time points.

The coefficients for the model are estimated using richly parameterized linear regressions, where a separate regression is used for each time \(t^*\), time into the future, \(t^* + t_s\), and length of history, \(t_h\). Thus, each model (2.2) has its own set of parameters; at this point, we do not propose to use information borrowing (shrinkage) either across or within models.

The full vector of covariates considered, \(W_{ikv}\), will include voxel-specific spatial information, demographic and treatment information. Spatial information incorporated in the model includes historical averages at each visit from the six “nearest” voxels (in 3D), meaning the voxels immediately above, below, and to each of the four sides of the voxel. Figure 2.2 provides a cartoon representation of the voxel (blue) with its nearest neighbors (red). The voxel in blue is the voxel where we are conducting the prediction, while the voxels in red are used to calculate the average “nearest-neighbors” intensity. This is done at each of the time points used from the voxel history. We also include the distance of the voxel to the boundary of the lesion, the distance being defined as how many voxels away is the nearest boundary of the lesion. Voxels on the lesion boundary are defined as having a distance of one.
Figure (2.2) Nearest neighbors: Visual representation of one voxel in a lesion. Blue cube represents the voxel specified, the six red cubes immediately surrounding the blue are the voxels used to calculate the average intensity of the “nearest” voxels in the spatial model.

Participant demographic and treatment information includes age (years), sex (M/F), disease subtype (relapsing-remitting or secondary-progressive), EDSS score, steroid use (Y/N), and treatment status (Y/N). Those who were treated received at least one of the following: Avonex, Betaseron, Daclizumab, and Rebif, in addition to experimental therapies. The model does not include time varying covariates other than the historical voxel information, but can be easily extended to incorporate additional time-varying covariates. The covariates such as age, EDSS score, steroid use, disease subtype and treatment can all change over the course of the study. We use the value of these covariates at the time of lesion incidence.
2.4 Results

2.4.1 Quantification and comparison of prediction accuracy

We first determine how much of the historical information to include by modifying two variables: (1) the “current” time $t^*$ that we predict from, and (2) the time $t^* - t_h$ of the earliest visit incorporated in the model. We compare two historical models (HM) at current times $t^* = 30$ and $t^* = 60$ (prediction is conducted after observing the lesion and following it either for one or two months). For each model we compare the average voxel root mean square error (RMSE) for various visit times where $t^* - t_h \in \{30, 60, 90, 120, 150, 210, 270\}$. We first focus on comparing models using only historical information for predicting lesion recovery based on RMSE. Table 2.2 shows the RMSE values for both historical models, including sub-models a-g. The amount of historical information used to fit the model is indicated in parentheses in the first column ($t^* - t_h, t^*$) and the time at which the FLAIR intensity is predicted is on the right hand side ($t = 60$ or $90$). For example, the first model (HM 1a) uses the voxel intensity at $t = -30, 0, 30$ as covariates to predict the voxel intensity at $t = 60$ and $90$, with the RMSE = 1.734 at $t = 60$ and RMSE = 1.749 at $t = 90$. As another example, HM 2c uses the voxel intensity at $t = -90, -60, -30, 0, 30$, and $60$ to predict the intensity at $t = 90$, having an average voxel RMSE of 1.272. Models HM 1a-g, only use historical information up through the first month after incidence ($t = 30$) to predict, HM 2a-g use information up through $t = 60$, which substantially decreases the RMSE from typical values around 1.74 to around 1.27. Additionally, as more historical information is included
into the past (going down the rows - HM 1a to HM 1g and HM 2a to HM 2g), the RMSE decreases, albeit only slightly for model 2.

**Table 2.2:** RMSE for first prediction point for models

<table>
<thead>
<tr>
<th>Model ((t^* - t_h, t^*))</th>
<th>Days after incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60</td>
</tr>
<tr>
<td>HM 1a (-30, 30)</td>
<td>1.734</td>
</tr>
<tr>
<td>HM 1b (-60, 30)</td>
<td>1.732</td>
</tr>
<tr>
<td>HM 1c (-90, 30)</td>
<td>1.316</td>
</tr>
<tr>
<td>HM 1d (-120, 30)</td>
<td>1.732</td>
</tr>
<tr>
<td>HM 1e (-150, 30)</td>
<td>1.731</td>
</tr>
<tr>
<td>HM 1f (-210, 30)</td>
<td>1.731</td>
</tr>
<tr>
<td>HM 1g (-270, 30)</td>
<td>1.731</td>
</tr>
<tr>
<td>HM 2a (-30, 60)</td>
<td>1.273</td>
</tr>
<tr>
<td>HM 2b (-60, 60)</td>
<td>1.272</td>
</tr>
<tr>
<td>HM 2c (-90, 60)</td>
<td>1.272</td>
</tr>
<tr>
<td>HM 2d (-120, 60)</td>
<td>1.272</td>
</tr>
<tr>
<td>HM 2e (-150, 60)</td>
<td>1.272</td>
</tr>
<tr>
<td>HM 2f (-210, 60)</td>
<td>1.271</td>
</tr>
<tr>
<td>HM 2g (-270, 60)</td>
<td>1.271</td>
</tr>
</tbody>
</table>

Note: RMSE = Root mean square error. \(^1\)The time in parentheses is the period of historical information used to fit model. Time listed is the number of days from lesion incidence. \(t^*\) = “current” time, or last time point used in model. \(t^* - t_h\) = first time point of historical data used in model.

As Table 2.2 shows, the historical models with the lowest error are those using historical scans out through \(t = 60\) (HM 2a-g). Adding one additional time point to HM 2a \((t = -60)\) only slightly decreases the RMSE (RMSE 1.273 to 1.272). It takes an additional five historical time points \((t = -90, -120, -150, -180, -210)\) to reduce the RMSE by a similar amount in model HM 2f (RMSE = 1.271).

Because HM 2b has slightly smaller RMSE than HM 2a, we use the historical trajectory defined by HM 2b, from \(t = -60\) to 60 in our historical models. Even though HM 2b has a slightly larger RMSE than HM 2f, it requires fewer visits for prediction. As more historical information is included with each successive model,
the number of lesions that have brain scans at least as far back as the "oldest" time point substantially decreases. Indeed, only 57\% of the 751 lesions have at least one visit 210 days prior to lesion incidence. In contrast 96\% of the lesions have scans with a visit earlier than 60 days before incidence. Thus, we choose to include historical information from visits 60 days before incidence to 60 days after (HM 2b).

Table 2.3: RMSE for each DLM

<table>
<thead>
<tr>
<th>Model (−60, 60)</th>
<th>90</th>
<th>120</th>
<th>150</th>
<th>180</th>
<th>210</th>
<th>240</th>
<th>270</th>
<th>300</th>
<th>330</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLM1 (history only)</td>
<td>1.27</td>
<td>1.44</td>
<td>1.44</td>
<td>1.50</td>
<td>1.54</td>
<td>1.66</td>
<td>1.69</td>
<td>1.67</td>
<td>1.67</td>
</tr>
<tr>
<td>DLM2 (demographic only)</td>
<td>2.60</td>
<td>2.54</td>
<td>2.34</td>
<td>2.36</td>
<td>2.36</td>
<td>2.34</td>
<td>2.36</td>
<td>2.40</td>
<td>2.36</td>
</tr>
<tr>
<td>DLM3 (spatial only)</td>
<td>1.43</td>
<td>1.56</td>
<td>1.53</td>
<td>1.57</td>
<td>1.62</td>
<td>1.74</td>
<td>1.76</td>
<td>1.74</td>
<td>1.74</td>
</tr>
<tr>
<td>DLM4 (history &amp; demographic)</td>
<td>1.27</td>
<td>1.43</td>
<td>1.44</td>
<td>1.49</td>
<td>1.53</td>
<td>1.65</td>
<td>1.69</td>
<td>1.66</td>
<td>1.66</td>
</tr>
<tr>
<td>DLM5 (history &amp; spatial)</td>
<td>1.25</td>
<td>1.42</td>
<td>1.43</td>
<td>1.48</td>
<td>1.52</td>
<td>1.65</td>
<td>1.68</td>
<td>1.66</td>
<td>1.66</td>
</tr>
<tr>
<td>DLM6 (spatial &amp; demographic)</td>
<td>1.42</td>
<td>1.55</td>
<td>1.53</td>
<td>1.56</td>
<td>1.60</td>
<td>1.72</td>
<td>1.75</td>
<td>1.72</td>
<td>1.72</td>
</tr>
<tr>
<td>DLM7 (all)</td>
<td>1.25</td>
<td>1.41</td>
<td>1.42</td>
<td>1.47</td>
<td>1.51</td>
<td>1.64</td>
<td>1.67</td>
<td>1.64</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Note: RMSE = Root mean square error; DLM = dynamic linear model.

1. The time \((t_1, t_2)\) is the period of historical information used for each of the models in the table. Time listed is defined as the number of days from lesion incidence. \(t_2 = 60\) is the “current” time, or last time point used in model. \(t_1 = −60\) is the first time point of historical data used in model.

We now focus on models that predict at the time \(t^* = 60\) (two months) after lesion incidence using historical data up to \(t_h = 120\) days before \(t^* = 60\), that is 60 days (two months) before lesion incidence. This is indicated in Table 2.3 as Model \((-60, 60)\). In this case, both \(t^*\) and \(t_h\) are fixed and we are interested in the effect of covariates on the prediction performance of these models at different time points after lesion incidence including \(t^* + t_s = 90, 120, 150, 180, 210, 240, 270, 300,\) and 330. Several dynamic linear models (DLM) are considered with different levels of covariate information: history (DLM1), demographic/treatment (DLM2), and spatial (DLM3), respectively. We also compared each combination
of these covariates, which created three additional models (DLM4-DLM6). The last model considered (DLM7) included all the covariates in models DLM1, DLM2, and DLM3. For each regression we estimate the prediction error using leave one out cross-validation (Ivanescu, Crainiceanu, and Checkley, 2017). Prediction of lesion trajectories is assessed using the voxel-level root mean square error (RMSE) at each future time point. We also visually examine the lesion predictions.

As expected, the farther into the future we predict, regardless of the other information included, the prediction gets progressively worse. Of the first three models the model using only historical information (DLM1) has, the smallest RMSE across all prediction times. The model using demographics information only (DLM2) performs the worst. It is unclear why for the demographics only model (DLM2) predictions get slightly better into the future (e.g., RMSE at $t = 90$ is 2.60 compared to 2.36 at $t = 330$). The model incorporating spatial information (DLM3) performs better than DLM2 and slightly worse than DLM1. This makes sense, as the voxel intensity of the neighboring voxels at time $t^* = 60$ are highly associated with the voxel intensity at the same time point, which is highly predictive of the future trajectory. The difference between DLM1 and DLM3 is more pronounced for earlier times than for later times. Specifically, at $t = 90$, the DLM1 RMSE=1.27 compared to DLM3 RMSE=1.43, a 13% difference, whereas at $t = 330$, DLM1 RMSE=1.67 compared to DLM3 RMSE=1.74, a 4% difference. Adding either demographic (DLM4) or spatial (DLM5) information to the model with just historical information (DLM1) improves predictions performance slightly (e.g., at $t = 120$, RMSE = 1.44 for DLM1 compared to 1.43 and 1.42 for DLM4 and DLM5, respectively). The
model with spatial and demographic information only (DLM6) performed slightly better than the model with just spatial information (DLM3), but not quite as well as the models including historical information. Adding demographic information (DLM7) to the model with just historical and demographic information (DLM4) improved prediction slightly (e.g., at $t = 120$, RMSE = 1.41 for DLM7 compared with 1.42 for DLM5). Hence, variables such as the FLAIR intensity of surrounding voxels, voxel distance to the lesion boundary, as well as treatment and demographic information appear to improve out of sample prediction. However, of all these variables, the most predictive is the voxel-specific historical intensity information.

2.4.2 Prediction Model Inference

2.4.2.1 Model coefficient estimators

We review interpretation of model coefficients from DLM1, DLM2, and DLM7 shown in Tables 2.4, 2.5, and 2.6 respectively. The majority of the coefficients were statistically significant regardless of the model. Thus, by a † to the left of the point estimator, we indicate those estimators that were not found to be statistically significant at $\alpha = 0.05$. Table 2.4 provides the point estimators for the parameters in DLM1 at each of the prediction times $t = 90, \ldots, 330$. Note that the last time point ($t = 60$) is by far the most predictive with point estimators, ranging from 0.814 at $t = 90$ to 0.503 at $t = 330$. This is intuitive for multiple reasons. First, it is the newest and most pertinent information concerning the current and future behavior of the lesion. Second, most lesion changes occur in the first 2-4 months.
after incidence Meier, Weiner, and Guttmann, 2007. Interestingly, the parameters before lesion incidence are also significant and positive, indicating that voxels with higher FLAIR intensity before incidence will tend to have higher intensities even after adjusting for the latest observed intensity. What is perhaps most unexpected is that the effect of the height of the spike in FLAIR intensity on the future intensities is very small after adjusting for the last observed and baseline intensities.

**Table 2.4:** Coefficient values ($\beta$) for DLM1 (historical model)

<table>
<thead>
<tr>
<th>Coefficients (history)</th>
<th>Prediction Times (future)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t=90</td>
</tr>
<tr>
<td>FLAIR, t=-60</td>
<td>0.050</td>
</tr>
<tr>
<td>t=-30</td>
<td>0.051</td>
</tr>
<tr>
<td>t=0</td>
<td>0.016</td>
</tr>
<tr>
<td>t=30</td>
<td>$^\dagger$0.002</td>
</tr>
<tr>
<td>t=60</td>
<td>0.814</td>
</tr>
</tbody>
</table>

1 All coefficients for $t = -60, -30, 0, 30, 60$ use historical FLAIR intensities to predict future FLAIR intensities at the prediction times shown at the top of the table (e.g. $t = 90, 120...$ etc.).

$^\dagger$p-value > 0.05, all other unmarked coefficients have p-values less than 0.05.

Table 2.5 provides results for the demographic and treatment effects model (DLM2) without using historical information. Just as in Table 2.4, we indicate by a $^\dagger$ to the left those estimators that were not found to be statistically significant at $\alpha = 0.05$. Positive coefficient values for age indicate that predicted intensities are slightly higher for older individuals in the immediate future (i.e., $t = 120, 150$). However, after $t = 180$ the association is reversed (note the negative coefficients), which is unexpected as remyelination is thought to be less efficient with age (Franklin, Zhao, and Sim, 2002). While this result is unexpected, we will investigate some of the potentially confounding effects. Table 2.5 indicate that males tend to have
lower intensities across all future time points (parameter estimates are negative), individuals with SPMS have higher intensities on average (almost all parameter estimates are positive). Moreover, study participants with higher EDSS scores have lower intensities, which is also slightly surprising, as EDSS scores are associated with higher burden of disease. Study participants who were taking steroids have higher intensities on average, after accounting for all other demographic and treatment variables. The treatment effects vary across time, which may be due to the definition of treatment, defined as treated/untreated at lesion incidence. The various treatment options can be highly heterogeneous across time points and individuals, and may be used more actively in study participants with specific symptoms or risk factors that are accounted for in our models.

Table 2.5: Coefficient values ($\gamma$) for DLM2 (demographic model)

<table>
<thead>
<tr>
<th>Coefficients (Demographic)</th>
<th>t=90</th>
<th>t=120</th>
<th>t=150</th>
<th>t=180</th>
<th>t=210</th>
<th>t=240</th>
<th>t=270</th>
<th>t=300</th>
<th>t=330</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.002</td>
<td>†0.002</td>
<td>0.012</td>
<td>-0.002</td>
<td>-0.009</td>
<td>-0.018</td>
<td>-0.006</td>
<td>-0.010</td>
<td>-0.008</td>
</tr>
<tr>
<td>Male</td>
<td>-0.564</td>
<td>-0.339</td>
<td>-0.257</td>
<td>-0.473</td>
<td>-0.432</td>
<td>-0.455</td>
<td>-0.278</td>
<td>-0.363</td>
<td>-0.135</td>
</tr>
<tr>
<td>SPMS</td>
<td>0.128</td>
<td>0.601</td>
<td>0.074</td>
<td>0.155</td>
<td>0.330</td>
<td>0.046</td>
<td>0.344</td>
<td>-0.290</td>
<td>0.192</td>
</tr>
<tr>
<td>EDSS</td>
<td>-0.101</td>
<td>-0.135</td>
<td>-0.122</td>
<td>-0.133</td>
<td>-0.164</td>
<td>-0.121</td>
<td>-0.146</td>
<td>-0.057</td>
<td>-0.073</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.104</td>
<td>-0.343</td>
<td>-0.360</td>
<td>*0.020</td>
<td>-0.120</td>
<td>0.106</td>
<td>0.041</td>
<td>-0.176</td>
<td>0.044</td>
</tr>
<tr>
<td>Steroid</td>
<td>0.125</td>
<td>0.261</td>
<td>0.775</td>
<td>0.304</td>
<td>0.332</td>
<td>0.837</td>
<td>0.782</td>
<td>0.567</td>
<td>0.281</td>
</tr>
</tbody>
</table>

1 Disease subtype of secondary progressive MS (SPMS), otherwise relapsing-remitting MS (RRMS); 2 EDSS = expanded disability status score. † p-value > 0.05, all other unmarked coefficients have p-values less than 0.05.

Table 2.6 provides the coefficient estimates for model DLM7, which adds historical and neighborhood (spatial) information to model DLM2. Most of the historical covariates in DLM7 (shown in the lowest 5 rows of 2.6) follow a similar pattern to those in DLM1 (shown in Table 2.4), especially for times $t = 0$ and $t = 60$. However,
the magnitude of the coefficients is substantially reduced, most likely because some of the information is incorporated in the average of the nearest neighboring voxels (note the middle rows labeled NN).

In addition, many of the coefficients from DLM2 covariates are substantially changed after accounting for history and neighborhood information. Indeed, at time 90 days (one month into the future) the point estimator for EDSS is now 0.045 compared to $-0.101$ in model DLM2. In general, EDSS coefficients are either substantially reduced in magnitude or have reversed signs. This makes much more sense biologically and indicates that higher EDSS either does not affect or adversely affects intensity trajectories after accounting for the known history of the voxel (i.e. individuals with higher EDSS scores have voxel intensities that are the same or higher on average). Indeed, a lesion that is more hyperintense in a study participant with a low EDSS is likely to stay more hyperintense than in a study participant with a higher EDSS. The effects of disease subtype are also substantially reduced in DLM7, indicating the strong confounding effect of the lesion-specific history. The effects of sex and steroid use are in the opposite direction in DLM7 compared to DLM2. The substantial change in coefficient values when the historical information is added is likely due to strong confounding between the historical information and EDSS, sex, disease subtype, and steroid use at the time of incidence, as well as to the low number of individuals in the study who had SPMS (6 study participants) and were taking steroids (3 at study start, though a few started taking them during follow-up).
### Table 2.6: Coefficient values ($\gamma, \beta$) for DLM7 (All)

<table>
<thead>
<tr>
<th>Coefficients (All)</th>
<th>Prediction Times (future)</th>
<th>t=90</th>
<th>t=120</th>
<th>t=150</th>
<th>t=180</th>
<th>t=210</th>
<th>t=240</th>
<th>t=270</th>
<th>t=300</th>
<th>t=330</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td>-0.005</td>
<td>0.002</td>
<td>0.005</td>
<td>$^\dagger$0.001</td>
<td>-0.010</td>
<td>-0.018</td>
<td>-0.012</td>
<td>-0.009</td>
<td>-0.003</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>-0.063</td>
<td>0.097</td>
<td>0.071</td>
<td>0.074</td>
<td>0.091</td>
<td>0.088</td>
<td>0.218</td>
<td>0.162</td>
<td>0.136</td>
</tr>
<tr>
<td>$^1$SPMS</td>
<td></td>
<td>-0.040</td>
<td>0.093</td>
<td>-0.158</td>
<td>-0.125</td>
<td>0.111</td>
<td>-0.203</td>
<td>$^\dagger$0.024</td>
<td>-0.208</td>
<td>-0.301</td>
</tr>
<tr>
<td>$^2$EDSS</td>
<td></td>
<td>0.045</td>
<td>-0.023</td>
<td>0.017</td>
<td>-0.044</td>
<td>-0.039</td>
<td>0.037</td>
<td>0.017</td>
<td>$^\dagger$0.007</td>
<td>$^\dagger$0.007</td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td>0.065</td>
<td>-0.152</td>
<td>-0.083</td>
<td>-0.026</td>
<td>-0.115</td>
<td>0.050</td>
<td>0.072</td>
<td>-0.082</td>
<td>0.052</td>
</tr>
<tr>
<td>Steroid</td>
<td></td>
<td>-0.223</td>
<td>-0.113</td>
<td>-0.069</td>
<td>-0.067</td>
<td>-0.277</td>
<td>-0.226</td>
<td>-0.141</td>
<td>-0.196</td>
<td>-0.231</td>
</tr>
</tbody>
</table>

$^3$DTB

| $^4$NN, t=-60     |                           | 0.146 | 0.115 | 0.157 | 0.225 | 0.268 | 0.166 | 0.217 | 0.251 | 0.131 |
| NN, t=-30         |                           | 0.058 | 0.108 | 0.027 | 0.228 | 0.097 | 0.121 | $^\dagger$0.026 | 0.270 | 0.254 |
| NN, t=0           |                           | 0.024 | -0.099 | $^\dagger$-0.004 | $^\dagger$0.001 | -0.042 | $^\dagger$0.018 | 0.029 | $^\dagger$0.007 | 0.060 |
| NN, t=30          |                           | -0.041 | -0.044 | -0.068 | -0.029 | -0.065 | -0.037 | -0.036 | -0.049 | -0.039 |
| NN, t=60          |                           | -0.188 | -0.233 | -0.197 | -0.241 | -0.187 | -0.219 | -0.207 | -0.175 | -0.203 |

$^5$NN, t=-60

| FLAIR, t=-60      |                           | $^\dagger$0.009 | 0.036 | 0.033 | $^\dagger$-0.018 | 0.022 | 0.038 | 0.048 | $^\dagger$-0.013 | $^\dagger$-0.005 |
| t=-30             |                           | 0.032 | 0.060 | 0.053 | 0.039 | 0.051 | 0.057 | 0.047 | 0.060 | 0.058 |
| t=0               |                           | 0.059 | 0.049 | 0.046 | 0.033 | 0.039 | 0.049 | 0.029 | 0.042 | 0.047 |
| t=30              |                           | 0.126 | 0.202 | 0.200 | 0.206 | 0.154 | 0.154 | 0.159 | 0.158 | 0.163 |
| t=60              |                           | 0.540 | 0.378 | 0.319 | 0.286 | 0.312 | 0.293 | 0.282 | 0.282 | 0.270 |

| FLAIR, t=-30      |                           | $^\dagger$0.009 | 0.036 | 0.033 | $^\dagger$-0.018 | 0.022 | 0.038 | 0.048 | $^\dagger$-0.013 | $^\dagger$-0.005 |
| t=-30             |                           | 0.032 | 0.060 | 0.053 | 0.039 | 0.051 | 0.057 | 0.047 | 0.060 | 0.058 |
| t=0               |                           | 0.059 | 0.049 | 0.046 | 0.033 | 0.039 | 0.049 | 0.029 | 0.042 | 0.047 |
| t=30              |                           | 0.126 | 0.202 | 0.200 | 0.206 | 0.154 | 0.154 | 0.159 | 0.158 | 0.163 |
| t=60              |                           | 0.540 | 0.378 | 0.319 | 0.286 | 0.312 | 0.293 | 0.282 | 0.282 | 0.270 |

1 Disease subtype of secondary progressive MS (SPMS), baseline is relapsing-remitting MS (RRMS); 2 EDSS = expanded disability status score; 3 DTB = distance to lesion boundary; 4 NN = average FLAIR intensity for 6 nearest voxels (NN = nearest neighbor); 5 All coefficients for $t = -60, -30, 0, 30, 60$ use historical FLAIR intensities to predict future FLAIR intensities at the prediction times shown at the top of the table (e.g. $t = 90, 120$...etc.).

$^\dagger$ p-value > 0.05, all other unmarked coefficients have p-values less than 0.05.

#### 2.4.2.2 Visualization of estimator effects on dynamic prediction

To better understand the effects of the models on prediction, Figure 2.3 displays the predictions for a voxel trajectory using models DLM1 and DLM7 when the history of intensities is perturbed at particular time points. The four panels on the top row correspond to DLM1 (A-D) and the four on the bottom row correspond to DLM7 (E-H). For example, the first panel in the top row (A) focuses on a voxel intensity trajectory (dashed blue line) that is modified by adding 1 (dashed red
line) or subtracting 1 (dashed black line) at time $t = -60$ and $t = -30$. The solid lines after $t = 90$ display the effect of predictions by keeping the same color coding. In this case, a higher (red) or lower (blue) baseline intensity correspond to higher and lower predicted intensities, respectively. The top row plots in the next two columns provide similar displays, but the blue trajectory is altered at time $t = 0$ (B) and $t = 30$ (C), respectively. In both cases predictions (solid lines) have barely distinguishable differences, conditional on the other historical information. However, last plot on the top row shows that a similar alteration at time $t = 60$ has a very large effect on predictions (D). This effect is one of the strongest we have quantified and is likely a signature for voxels within the same lesion that have more extreme recovery behaviors (either very fast or very slow).

The four panels on the bottom row (E-H) are very similar to the ones on the top, except that they also account for demographic and treatment effects as well as neighborhood information. Each plot from DLM7 corresponds to the plot directly above it. The results are qualitatively similar, though the difference between predictions is substantially attenuated. This is probably due to the fact that models take into account voxel neighborhood information, which is highly correlated with the voxel intensity. Similar results can be observed for the other panels, with the exception of the third column (C, G). In this case the prediction lines are farther apart for model DLM7 (G) than for model DLM1 (C).

Figure 2.4 provides a comparison of prediction effects of models DLM2 (four panels on the top) and DLM7 (four panels on the bottom) when only one covariate
Figure (2.3) Historical Variables: Demonstration and comparison of historical effects for DLM1 and DLM7 models. The four plots on the top row are from DLM1 (A-D) and those on the bottom row are from DLM7 (E-H). The dashed line in each plot ($-60 \leq t \leq 60$) is an example of the trajectory for one voxel used to predict future intensities in each model. The solid lines in each plot show the predictions for each model ($90 \leq t \leq 330$). The corresponding dotted lines are the 95% confidence intervals for the mean predicted values. Each color corresponds to a different “hypothetical individual/voxel” for the explanatory variables. For example, in the first plot of each row for both DLM1 and DLM7, the only explanatory variables that are changed are at $t = -60$ and $-30$. The red dashed curve has $(+1)$ higher intensity at both time points, the black curve has $(-1)$ lower for both, and the blue curve is in the middle. The corresponding predictions are shown in the same color. In the third plot of each row for DLM1 (C) and DLM7 (G), the only difference is at $t = 30$ where the differences are again $(+1)$ and $(-1)$ for the black and red curves. The remaining plots are similar and show predictions when the historical intensities are perturbed at specified time points.

is changed while the others are kept constant. This, of course, should not be over-interpreted as the correlation between variables is quite high (e.g., among historical intensities and between voxel and neighboring voxel intensities). The thick dashed line in each panel ($-60 \leq t \leq 60$) is an example of the trajectory for one voxel used to predict future intensities in each model. The solid lines in each plot show
the predictions for each model \((90 \leq t \leq 330)\), while the dotted lines are the 95% confidence intervals. Each color corresponds to a predicted trajectory for a different value of the explanatory variables, while keeping all other covariates fixed.

Figure (2.4) Demographic Variables: Demonstration and comparison of demographic effects for DLM2 and DLM7 models. The four plots on the top row are from DLM2 (A-D) and the four on the bottom row are from DLM7 (E-H). The dashed line in each plot \((-60 \leq t \leq 60)\) is an example of the trajectory for one voxel used to predict future intensities in each model. The solid lines in each plot show the predictions for each model \((90 \leq t \leq 330)\). The corresponding dotted lines are the 95% confidence intervals for the mean predicted values. Each color corresponds to a different “hypothetical individual/voxel” for the explanatory variables. The hypothetical individuals represented by each curve have a unique value for one covariate, (e.g. age, sex) and are identical for all other covariates. Covariates examined include age, sex, EDSS score, and disease subtype (RRMS or SPMS). For example, in the first plot on each row for both DLM2 and DLM7, the only covariate that is changed is age. The three curves compare predicted trajectories (solid lines) for individuals that are 25, 35, and 45 years of age (red, blue, black curves respectively) and have all other characteristics in common. The subsequent plots are similar, but vary according to the other three covariates.
The first panels in the each row (A, E) explore the effect of age on predicted trajectories: Age = 25 (red), Age = 35 (blue), and Age = 45 (black). Model DLM2 predicts that for younger study participants the voxel recovers faster in the first couple of months and slower after that. This could be due to a floor effect in recovery, but more needs to be done to explain this unexpected finding. The corresponding results for DLM7 are slightly different with no difference between younger and older study participants in the first two months followed by slower recovery for younger study participants 150 days after lesion incidence. Because DLM7 incorporates both the history up to day 90 after lesion incidence, this result suggests that recovery for younger study participants may happen earlier in the lesion development process. However, this “faster recovery and floor effect” hypothesis requires further investigation.

A puzzling result was obtained for EDSS scores using model DLM2 (B). Indeed, higher EDSS score (more severe disease) is associated with lower voxel intensities after day 90. The puzzle is partially solved by model DLM7 (panel four in the first row), which indicates that these differences are confounded by the history of the intensities up to day 90 after lesion occurrence. This suggests that the voxel history may be associated with EDSS scores. The sex effect using model DLM2 (first panel in the second row) suggests that women have a higher average intensity in lesion voxels. However, model DLM7 reverses the effect (see G) indicating that conditional on the historical information up to day 90 after lesion incidence women are expected to have a lower average intensity. A similar story seems to hold for disease subtype (see panels D and H).
Figure 2.5 is similar to Figure 2.4, though it focuses on whether or not treatment and steroid were being used at the time of lesion incidence. Models DLM2 and DLM7 (panels A and C) agree that the average voxel intensity after day 90 from lesion incidence is lower for the treated group, irrespective of whether or not one conditions on the history of the trajectory. However, in the case of steroid use the two models provide contrasting results. DLM2, which does not account for history, predicts that individuals who are using steroids at baseline will have a higher intensity (panel B). However, once one accounts for the voxel intensity up to day 90, the effect is changed (panel D). This result indicates a potentially strong confounding effect between the history of the lesion and the use of steroids. We conclude that the parameter estimates in this context do not have a causal interpretation and should only be interpreted in the context of an in-depth exploratory analysis. The main goal of our paper is dynamic prediction.

Figure (2.5) Treatment Variables: Treatment effects for DLM2 and DLM7 models. The two plots on the left are from DLM2 (A,B) and the two on the right are from DLM7 (C, D). The dashed line in each plot ($-60 \leq t \leq 60$) is an example of the trajectory for one voxel used to predict future intensities in each model. Similar to Figure 2.4, we compare voxels where the individual was undergoing treatment (A, C) or taking steroids (B, D) at the time of incidence (blue is voxels that were taking treatment or steroid, red is voxels that were not).


2.5 Visualization and software

2.5.1 Results for two lesions

To better visualize the performance of prediction approaches, panels A and B in Figure 2.6 display the observed (black) and predicted (red) data for all voxels from a slice of lesion 55. In panel A we emphasize (note the thicker lines) one voxel with RMSE=0.322, while in Panel B we emphasize a second voxel (again, note the thicker lines) with RMSE=0.416 using model DLM7. We show these two voxels because they both are in the lowest 10% RMSE values among all voxels and lesions. Panels C and D show the observed and predicted data at every voxel in the slice, respectively for each visit included in the model. Voxels A and B (identified by small arrows in panels C and D) indicate the locations of the voxels displayed in panels A and B. Each square in panels C and D represents an individual voxel and is colored according to either the actual (C) or predicted (D) normalized FLAIR intensity. The color legend next to Figure 2.6-D is the same for both plots, with blue corresponding to lower values, red to higher values, and white centered at 2.5. This coloring scheme shows voxels that have returned to their baseline mean levels in blue (within $\sim 2$ standard deviations of baseline mean) and voxels that have not returned in red (normalized values are still greater than 3 standard deviations). For most of the lesion slice, the predicted outcome shown in Figure 2.6-D appears to agree with the actual intensities on the row above.

We make a similar plot for two voxels from a slice of lesion 19, but with prediction RMSE in the highest decile of the RMSE distribution across all voxels.
Figure (2.6) Prediction Example 1: Example of lesion slice comparing predicted values from DLM7 to original interpolated values. A) and B) show original data in black/gray and predicted values in red, each curve is one voxel from the lesion slice. A) highlights a single voxel in lowest decile of RMSE (0.322) and B) highlights another voxel in the same lesion with RMSE also in the lowest decile (0.416). C) and D) show the lesion slices at each visit included in the model, historical and future. C) shows the actual normalized intensities and D) shows the predicted values. The voxels are colored according to intensity of the voxel from low (blue) to high (red). The voxels highlighted in A) and B) are indicated with arrows in both rows and labeled in the plot of the lesion at the first visit in the row.

Figure 2.7 has an identical setup as Figure 2.6, but displays the intensity trajectories of two different voxels from a different lesion (thick black lines). The two voxels highlighted in panels A and B (and marked with arrows in panels C and D) have RMSE of 3.035 and 2.308 in model DLM7, respectively. The poor prediction performance extends to the other voxels in this lesion (shown in solid gray in panels
Figure (2.7) Prediction Example 2: Example of lesion slice comparing predicted values from DLM7 to original interpolated values. A) and B) show original data in black/gray and predicted values in red, each curve is one voxel from the lesion slice. A) highlights a single voxel in highest decile of RMSE (3.035) and B) highlights another voxel in the same lesion with RMSE also in the lowest decile (2.308). C) and D) show the lesion slices at each visit included in the model, historical and future. C) shows the actual normalized intensities and D) shows the predicted values. The voxels are colored according to intensity of the voxel from low (blue) to high (red). The voxels highlighted in A) and B) are indicated with arrows in both rows and labeled in the plot of the lesion at the first visit in the row.

A and B). One could argue that the bold red (predicted) curves from panels A and B appear to be reasonable predictions given the data up to $t = 60$. However, in this case something unexpected happens. The data does not level off around zero, as one would expect when voxel intensities return to their pre-incidence distribution. Instead, the normalized voxel intensities continue to drift into more negative values.
close to −4. This indicates that voxels became hyperintense immediately after the lesion was formed (which is expected) but then became hypointense relative to their own baseline levels (which is unexpected). The source of this observed behavior is not likely to be biological and we believe that the predictions in this case look more reasonable than the observed data. Indeed, there could be several possible explanations for these unusual observed voxel intensity trajectories. First, data may be mis-registered towards the end of the longitudinal observations, which would incorrectly assign a darker area after time $t = 90$ to a lesion voxel. Second, the baseline variability of voxel intensities may be too small and the number of observations may not be enough to capture the true variability of the voxels. This could exacerbate the negative values of the trajectory by dividing by an unusually small standard deviation. Third, the NAWM normalization procedure may be biased in this particular longitudinal sequence. Regardless of the reason, this example highlights just how much variability can be present in the FLAIR intensities from visit to visit.

This second example is an illustration of what goes right with the model. Indeed, despite the very large RMSEs, the predictions are actually reasonable. The reason is that the model is borrowing strength from the other lesions and deduces that there is overwhelming evidence that, “if a lesion is detected at time zero and the normalized intensity recovers close to its pre-incidence level, then the future intensity trajectory is likely to slowly decrease or stay the same.” Thus, some of the large RMSEs obtained for some of the lesions may indicate problems with data and not with the prediction models. Nevertheless, the visual examination of the results
provides an excellent way to explore both the data and the prediction results. In the next section we describe a fast visualization tool using Shiny in R (R Core Team, 2018) for all data and prediction models.

### 2.5.2 Visualization via Shiny in R

Shiny applications in R (R Core Team, 2018) provide fast, interactive, and intuitive methods for exploring and visualizing data. We created a Shiny application to visualize lesion behavior over the course of follow-up using a time-series, slice, and 3D representation of the longitudinal lesion intensity data. The application can be found online at [https://hotavocado.shinyapps.io/lesionView2/](https://hotavocado.shinyapps.io/lesionView2/) and for reference here, three screenshots of the interactive app are shown in Figure 2.8.

There are three tabs in the application, shown at the top of each screenshots in Figure 2.8: Lesion, Slice, and 3-D. The first tab, “Lesion”, shown in the upper left panel of Figure 2.8, shows an entire lesion at each visit during follow-up on both the original scale (normalized using the NAWM) and on the new scale (original intensities normalized by pre-incidence intensity levels). The columns are organized by visit and the rows are two-dimensional slices of the lesion. This view provides an overall idea of what is happening to the entire lesion over time. For example, we can clearly see the abrupt change from blue to red in the slices’ coloring, indicating the formation of the lesion. This coloring can be adjusted in the upper right hand corner of the panel.

The second tab “Slice”, shown in the upper right screenshot in Figure 2.8, displays more details for a specified lesion slice. The slice image, an extraction of
Figure (2.8) Shiny Application: Screenshots from Shiny application for lesion visualization. There are three tabs, the first tab on the left shows a FLAIR intensities in a lesion through each slice of the lesion, using both the original data standardized by the NAWM and the data normalized by the pre-incidence intensities. The second tab focuses on individual slices, showing the trajectories from all voxels in the slice as well as the predicted values from the model. On the bottom of the screenshot, a series of one lesion slice is shown for each visit with the original, normalized, and predicted intensities. The third panel (bottom row) shows the lesion in three dimensions, viewing one visit at a time over the course of individual follow-up.

one row from the plot from the tab described in the previous paragraph (Lesion), is shown at the bottom of the screen for each visit. The trajectories of the voxels from the chosen slice, including both original and pre-incidence normalized scales, are
shown in the spaghetti plots in the upper right corner of the second screenshot; each line is one voxel from the indicated slice. The corresponding predicted trajectories from DLM7 are shown as curves in the third panel. Not shown in this screenshot, the application also contains predictions for the specific slice in the same format and coloring scheme as the images slices. After getting an overall lesion view in the first tab, the second tab provides a more detailed look at the slice or voxel level. This allows the visual comparison of the predicted values and intensities, as shown, for example, in Figure 2.6. The color scale and lesion slice are controlled using the panel on the left hand side of the screen.

The third tab, shown in the third screenshot at the bottom of Figure 2.8, provides the 3D view of the entire lesion. Currently, it is only set up to view one visit at a time, but the visit number can be changed using the slider bar in the bottom right corner. The lesion can be rotated to view from various angles by clicking and dragging it around with the mouse. The lesion being viewed can also be changed within each tab. This application provides a way to view and examine each lesion one at a time quickly and efficiently, without having to restart or rerun code to view a different lesion. The interactive nature of the application and its features have provided us with a much quicker and complete understanding of lesion recovery and model performance than would have been possible plotting each lesion individually through traditional methods.
2.5.3 R code and vignette

In addition to the shiny application, we have also created a vignette, which provides step-by-step instructions for the process of transforming data from the format in Table 2.1 to the format the analytic format used for dynamic prediction models. The vignette also includes code for visualization of prediction results, as shown in Figure 2.6 and 2.7. In this section, we provide an outline of the vignette and include some R code (R Core Team, 2018) for fitting the dynamic linear models.

1. Using the full data, isolate the visits before lesion incidence. Using these visits, calculate the voxel-specific mean and standard deviation, as well as the average voxel standard deviation for each lesion.

2. Calculate the voxel-specific stabilized variance by averaging the voxel- and lesion-specific average variances; see Section 2.3.1. Calculate z-scores for all voxels by subtracting the voxel-specific pre-incidence mean and dividing by the stabilized standard deviation.

3. Split data into two parts: 1) historical data, including all visits where $t \leq 60$, and 2) future data, including all data from visits where $t > 60$. Interpolate normalized FLAIR intensities on a 30-day equally spaced grid for the historical and future data separately, as described in Section 2.3.2.

   a. For historical data, use linear interpolation to calculate normalized intensities at 30 day intervals
b. For future data, use flat interpolation based on the average of all visits within a 30 day window, centered at \( t = 90, 120, \ldots, 330 \).

4. For each lesion voxel, identify the six nearest voxels and calculate their average intensity at each historical time point. Merge the historical and future data and together with the distance to the boundary (DTB) for each voxel.

5. Next, merge voxel level data (produced using steps 1-4) with the demographic and treatment data for each individual and lesion.

A small subset of the resulting data (model_dat) is shown below with R code snippets used in the subsequent steps to fit the dynamic linear models. The data now has lesion and subject information (id, lesionid, xind, yind, zind, xid), demographic information (age_y–steroid), spatial information (dtb and variables starting with knn), as well as historical and future intensities (i.e., \(-60, -30, \ldots, 330\)). The NA values at \( t = 180 \) and 330 indicate there were no visits for this individual within 15 days of 180 and 330.

```r
library(tidyverse)
file_pre<-"data_working_directory/"
model_dat <- readRDS(paste0(file_pre,"datafile.rds"))
pred_times_future<-as.character(c(90,120,150,180,210,240,270,300,330))
head(model_dat)
```

<table>
<thead>
<tr>
<th>id</th>
<th>lesionid</th>
<th>xind</th>
<th>yind</th>
<th>zind</th>
<th>xid</th>
<th>dtb</th>
<th>age_y</th>
<th>sex</th>
<th>subtype</th>
<th>edss</th>
<th>trt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>119</td>
<td>99</td>
<td>1</td>
<td>49.1</td>
<td>F</td>
<td>RRMS</td>
<td>2.5</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>123</td>
<td>120</td>
<td>99</td>
<td>2</td>
<td>49.1</td>
<td>F</td>
<td>RRMS</td>
<td>2.5</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>124</td>
<td>120</td>
<td>99</td>
<td>3</td>
<td>49.1</td>
<td>F</td>
<td>RRMS</td>
<td>2.5</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>125</td>
<td>120</td>
<td>99</td>
<td>4</td>
<td>49.1</td>
<td>F</td>
<td>RRMS</td>
<td>2.5</td>
<td>Yes</td>
</tr>
</tbody>
</table>

49
6. Prepare data for model fitting and prediction. The voxel being predicted is removed from the model (row $i$). This process is repeated for each voxel and prediction time point and is shown in the R code below. Since there are over 187,000 voxels, the process is parallelized, running groups of 1,000 voxels at a time, each on a different node. Prediction and model fitting takes approximately 30-40 minutes for each group of 1,000.

```r
# Set up: remove variables not needed
Y = select(model_dat, -xid, -id, -lesionid, -xind, -yind, -zind)
x = nrow(Y) # or however many is done at a time - here 1,000
ypred <- matrix(nrow = x, ncol = length(pred_times_future))

# This will be repeated for every voxel (parallelization not shown here)
# for (i in 1:x) {
  for (j in 1:length(pred_times_future)) {
    # Organize model data (remove observation being predicted, choose prediction time)
    data.mod1a <- Y[-i,] %>% select(one_of(pred_times_future[j]))
  }
}
data.mod1b <- Y[-i,] %>% select(-one_of(pred_times_future))
data.mod1 <- cbind(data.mod1a, data.mod1b)
names(data.mod1)[1] <- "y.resp"

After this step, the data should look somewhat like the following subset, depending on what terms are included in the model. The first column is the prediction time (response), followed by DTB, then demographic, spatial, and historical variables used to fit the model.

head(data.mod1)
y.resp dtb age_y sex subtype edss trt steroid knn-60 knn-30 knn0 knn30 knn60 -60 -30 0 30 60
1 NA 1 49.1 F RRMS 2.5 Yes 0 0.628 1.82 4.85 3.60 2.88 -0.3177 2.286 4.14 1.97 2.71
2 NA 1 49.1 F RRMS 2.5 Yes 0 0.394 1.56 5.67 3.83 3.03 0.0555 1.371 6.33 3.46 3.24
3 NA 1 49.1 F RRMS 2.5 Yes 0 0.868 1.55 4.69 3.39 2.68 0.4934 1.689 3.50 3.10 2.79
4 NA 1 49.1 F RRMS 2.5 Yes 0 0.671 1.48 4.59 2.94 2.60 1.2352 1.310 3.45 3.60 3.01
5 NA 1 49.1 F RRMS 2.5 Yes 0 0.565 1.69 5.19 3.05 2.64 -0.0646 0.564 4.08 2.41 2.56
7 NA 1 49.1 F RRMS 2.5 Yes 0 0.716 1.68 4.82 3.13 2.48 1.1869 0.910 3.02 2.56 1.80

7. Fit the dynamic linear model for future time \( j \) (see step 6 for where the model fitting is repeated for each future time) using all covariates included in the data (data.mod1).

# Fit model
fit.mod1 <- lm(y.mod1 ~ ., data = data.mod1)

8. Predict FLAIR intensity at the prediction time using model information for the current voxel; see R code below (after step 9).
9. Repeat model fitting (steps 6-8) for each voxel and each time ($t = 90, \ldots, 330$).

```r
# Predict intensity for observation left out
new.data.mod1 <- Y[i,] ## select(-one_of(pred_times_future))
ypred[i,j] <- predict(fit.mod1, newdata = new.data.mod1)
```

# Close for loop (all time points)
}

# Add column names to prediction results
colnames(ypred) <- c("t90", "t120", "t150", "t180", "t210", "t240", "t270", "t300", "t330")

The resulting predictions should look something like the subset below, where each row contains the predictions for each individual voxel at all time points.

```r
head(ypred)
t90 t120 t150 t180 t210 t240 t270 t300 t330
[1,] 2.152 1.640 1.616 1.530 1.266 1.245 1.188 1.374 1.529
[2,] 2.682 2.177 2.096 1.944 1.650 1.645 1.570 1.688 1.854
[3,] 2.270 1.888 1.822 1.758 1.444 1.427 1.364 1.551 1.699
[4,] 2.492 2.137 2.057 1.930 1.633 1.626 1.586 1.664 1.828
[5,] 2.070 1.600 1.572 1.521 1.203 1.202 1.165 1.285 1.460
[6,] 1.850 1.464 1.425 1.381 1.061 1.056 1.040 1.112 1.274
```

The vignette concludes with a section containing code to view specific lesion slices and voxels to compare predicted values to the actual responses at each time point, similar to the plots in Figures 2.6 and 2.7.

### 2.6 Discussion

In this paper, we developed methods for dynamic prediction of the future trajectory of voxel-level intensities in MS lesions. To do that, we introduced a normalization method based on z-scoring relative to the voxel-specific baseline mean intensity.
and a stabilized standard deviation. Prediction performance was assessed and compared using cross-validated RMSE. Results indicate that: (1) prediction accuracy improves substantially when historical information is available, with the last observation being the strongest predictor; (2) good prediction can be obtained using monthly scans as far as one year into the future; (3) models including only spatial covariates or demographic and treatment information substantially underperformed models that contained historical information; (4) adding demographic, treatment or spatial components to historical data models only slightly improved prediction; and (5) prediction performance decreased the further the prediction time was from the time when prediction is conducted.

The dynamic linear models performed well in spite of the irregular spacing of the visits and the substantial amount of noise in the data. Moreover they are fast, relatively easy to implement, and easy to explain. We have also considered functional regression approaches together with smoothing to borrow strength across models. However, we currently do not have a working functional regression solution because of the very large size of our data set, which includes over 187,000 voxels.

One limiting factor of this study is that the exact date of lesion formation is unknown. Indeed, a lesion is only observed when an sMRI scan is conducted, but the lesion could have occurred at any time between the previous and current scan. Currently, the voxel intensity trajectories are treated as if they all start at \( t = 0 \), the moment when the lesion was observed. Ideally, we would like the time to be aligned when the lesion actually occurs, not when it is observed. Accounting for
measurement error in the alignment of lesion incidence may improve the accuracy of the prediction approaches.

Dynamic prediction using cross-validated, voxel-specific models is an important early step in understanding lesion behaviour and predicting how long it takes for lesions to return to baseline levels, or if they return at all. We contend that we introduced a reasonable way to account for the extreme variation and irregularity in timing between visits given the complexity and size of the dataset. We also introduced a novel approach to normalizing FLAIR intensities, which allows us to compare voxel intensities, visually and quantitatively, after lesion incidence to typical voxel-specific baseline intensity levels. This is a major methodological contribution that likely outweighs the specific prediction method that we or others may propose. Additionally, being able to visualize and understand the probable trajectory of recovery from a given point forward in time has the potential to be an extremely useful tool for clinicians and researchers moving forward. With further model expansion and developments, these tools and methods could be used to help advise identify early patterns of lesion development and inform personalized treatment decisions.

The most important component of our approach is that we provide detailed visualization tools via a Shiny app together with an R package and detailed vignette. These tools allow fast, in-depth analysis of the data and provide insights into what data may be too heterogeneous or have unusual behavior. Moreover, starting with these reproducible components, future research with this unique data should be relatively easy to continue.
References


Chapter 3

Fragmentation as a Novel Measure of Stability in Normalized Trajectories of Mood and Attention Measured by Ecological Momentary Assessment

3.1 Introduction

There has been rapid growth in the use of mobile technologies that collect intensive repeated measures of activity, mood, diet, exercise, sleep, and other domains in both daily life and research studies. In the field of psychology, these measures have long been employed through methods known by a variety of names, such as ecological momentary assessment (EMA), ambulatory assessment, and the experience sampling method (ESM). They have the potential to explore the dynamic relationships between an individual’s experiences, environment, and responses, even over time periods as short as a day (Ebner-Priemer and Trull, 2009).

A wide variety of measures have been used to quantify within-participant variability and stability in EMA applications, and each measure highlights unique features that are important in different contexts (Nesselroade and Ram, 2004). For
example, one of the most common measures used to quantify variability is the participant-specific standard deviation. The residual individual standard deviation is a slight variation of this, which uses the time and participant specific expected value, $\mu_{ij}$, in place of the typical participant-specific mean, $\mu_i$ (Stamps, Briffa, and Biro, 2012). The most common approach to estimate $\mu_{ij}$ is by fitting multilevel models with participant-level random effects. Ferrer, Steele, and Hsieh, 2012 characterized stability with periods of relative constancy, defined as periods of time where the maximum distance is less than some threshold. The intraclass correlation coefficient from a multilevel model has also been employed to determine if there were differences in within-individual variability in positive and negative affect between two age groups (Röcke, Li, and Smith, 2009).

As momentary data are temporally dependent, it is important that assessments of variability and stability account for the autocorrelation inherent in the data. One of the most common measures of stability is therefore the mean squared successive difference (MSSD), which is the average squared difference between each pair of subsequent observations within a single participant (Ebner-Priemer et al., 2009; Jahng, Wood, and Trull, 2009; Neumann et al., 1941). MSSD or its variants (i.e., root mean squared successive difference) have been used to show that individuals with bipolar disorder or major depression have higher affective instability compared with control group participants (Ebner-Priemer et al., 2007; Koval et al., 2013; Stein, 1996; Tsanas et al., 2016; Woyshville et al., 1999). However, as important differences may be observed across disorder type and according to the alternative metrics for
quantifying affective variation or instability (Houben, Van Den Noortgate, and Kuppens, 2015), direct comparisons within the same investigation are needed.

Fragmentation analysis was originally developed to characterize patterns of sedentary time accumulation in objectively measured physical activity. Specifically, fragmentation analysis has been used to quantify either the frequency of transitions between active and sedentary states or the duration of bouts of sedentary and active behaviors (Di et al., 2017). In this paper, we describe the application of these measures in a study using EMA to assess the stability of emotional states in subgroups of people with mood disorders including: Bipolar I (BPI), Bipolar II (BPII), and major depressive disorder (MDD), compared with controls with no lifetime history of these disorders.

Previous analyses of EMA data from the current study examined variability and changes in emotional states as a function of the valence of daily life events (Lamers et al., 2018). Here, we introduce the use of between-state transition probabilities, as measures of fragmentation. Our primary objective is to provide an illustration of how fragmentation analysis may offer additional insight into the stability of these indicators of emotional states and attention beyond that obtained by traditional measures such as the participant-specific standard deviation and MSSD. Our secondary aim is to investigate whether instability of mood and attention differs between people with a history of bipolar disorder or major depression compared with unaffected controls.
3.2 Methods

3.2.1 Data

Data from the mobile technology component of the National Institute of Mental Health (NIMH) Family Study of Affective Spectrum Disorders, a large community-based sample in the Washington DC metropolitan area, were used in these analyses (Merikangas et al., 2013). Participants were recruited through mail contacts, screening a sample of more than 11,000 households within 50 miles of Washington DC. Twenty percent of households that did not respond to any of the mailings were contacted by study recruiters to encourage their participation. Inclusion criteria were an ability to speak English and availability and consent to contact at least two living first-degree relatives.

3.2.2 Procedures

The procedures and measures for the NIMH Family Study are reported in Merikangas et al., 2013. Briefly, probands and their relatives were directly interviewed by a team of experienced clinicians (licensed psychologists and a board-certified psychiatrists) using a structured diagnostic interview, the NIMH Family Study Diagnostic Interview, which is adapted from Schedule for Affective Disorders and Schizophrenia/Diagnostic Interview for Genetic Studies. Structured family history reports were obtained from probands and relatives about all family members. Direct interview and family history data were combined to characterize mood and other mental disorders based on the standard best estimate diagnoses for family
study data (Leckman et al., 1982). In this investigation, we compare those with a history of MDD, BPI, and BPII with controls who have no history of any lifetime mental disorder.

3.2.3 Ecological Momentary Assessment (EMA)

A total of 371 people from the NIMH Family Study who enrolled in the National Institutes of Health Clinical Center evaluation completed an electronic survey using a personal digital assistant (Tungsten E2 PDA, Sunnyvale, CA) four times per day over a 2-week period (rendering a maximum of 56 assessments per individual). Signaling schedules were adapted to accommodate for the individual’s usual sleep/wake cycles, occurring at fixed intervals for each participant but with signal times that were randomized across participants. The signal occurred once within four daily time epochs of three to four hours in length, and with a minimal spacing between signals of two hours. Participants completed the 2-week EMA assessment period at various times across the entire year, thereby avoiding the over- and undersampling of specific experiences associated with time-dependent events (i.e. seasons, specific holidays). All participants received training from a research assistant on how to use the PDA and how to complete the surveys. Research assistants called each participant after the start of the EMA measurements to verify that they understood study procedures and that the PDA device was functioning properly.

EMA assessments included questions concerning mood and cognitive states, daily event occurrences, and diverse additional variables assessing behavior or
experiences (e.g., activities, appetite, food intake, sleep, somatic complaints). For
the purpose of the present investigation, one self-reported mood state and one self-
reported cognitive state were selected for analysis. Based on the mood circumplex
(Larsen and Diener, 1992), we examined patterns of responses to seven-point
Likert scales assessing current mood ranging from very happy (1) to very sad (7) and
perceived attention levels ranging from very distracted (1) to very focused (7),
in the major mood disorder subgroups of BPI, BPII, and MDD, compared with
controls without a lifetime history of mood disorders. Table 3.1 summarizes the
demographic characteristics for each of the mood disorder and control groups,
showing that the sample was predominantly female (i.e., 60.9%), White (72.7%),
and the average age in years was 42.6.

Table 3.1: Demographic Information Across Mood Disorder Subgroups

<table>
<thead>
<tr>
<th>Category</th>
<th>Unit</th>
<th>BPI</th>
<th>BPII</th>
<th>MDD</th>
<th>Control</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age - Mean (SD)</td>
<td>Years</td>
<td>41.1 (14.2)</td>
<td>36.9 (18.3)</td>
<td>45.1 (17.7)</td>
<td>42.5 (22.8)</td>
<td>42.6 (19.9)</td>
</tr>
<tr>
<td>Sex - N (%)</td>
<td>Male</td>
<td>13</td>
<td>17</td>
<td>39</td>
<td>76</td>
<td>145 (39.1%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>22</td>
<td>27</td>
<td>94</td>
<td>83</td>
<td>226 (60.9%)</td>
</tr>
<tr>
<td>Race - N (%)</td>
<td>White</td>
<td>29</td>
<td>30</td>
<td>102</td>
<td>108</td>
<td>269 (72.5%)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>6</td>
<td>14</td>
<td>31</td>
<td>51</td>
<td>102 (27.5%)</td>
</tr>
<tr>
<td>Total - N (%)</td>
<td></td>
<td>35 (9.4%)</td>
<td>44 (11.9%)</td>
<td>133 (35.8%)</td>
<td>159 (42.9%)</td>
<td>371</td>
</tr>
</tbody>
</table>

Note: BPI = Bipolar I; BPII = Bipolar II; MDD = major depressive disorder.

3.2.4 Participant-Level Normalization

We first normalized participant-specific EMA trajectories using a two-step ap-
proach. At step one, the selected EMA scores were normalized at the participant
level, subtracting the participant-specific mean and dividing by the participant-specific standard deviation. The normalization facilitates the comparison of stability even for two participants with vastly different ranges for their EMA responses. For example, Figure 3.1 presents the raw and transformed data for two participants. The top two plots show the raw data and the middle two plots show the normalized data for the two participants. For the participant 2 in the right column of Figure 3.1, the range of responses is relatively narrow, from about two to four, while the participant on the right side has a much larger range of responses, from one to six. Although the spread of the normalized responses for the two participants is approximately the same, the normalization helps account for the participant-level differences.

At step two, the normalized EMA scores are dichotomized according to whether the score is inside or outside the participant-specific standard range. We define being inside the standard range when the participant is within one standard deviation of their mean, and define being outside the standard range as when the EMA score is more than one standard deviation from the mean in either direction using the participant-specific mean and standard deviation. This second transformation step can be seen comparing the lower two plots to the respective middle plots of Figure 3.1. The dotted lines in the middle plots show the plus/minus one standard deviation range for each participant.

Standardizing the data before dichotomization allows us to more accurately compare the time spent inside and outside the standard range across participants. The second participant (right column) only has four unique responses, so they are
Figure (3.1) EMA: Data Example and Transformation: Raw, normalized, and dichotomized transformations of exemplary Ecological Momentary Assessment (EMA) attention scores from two participants. With respect to the raw data, low levels (1 to 3) correspond to a distracted state and high levels (5 to 7) correspond to a focused state. For the binarized data, ISR (inside standard range) corresponds to normalized assessments that are within one standard deviation of the participant mean, whereas OSR (outside standard range) corresponds to normalized assessments that are greater than one standard deviation away from the mean.

outside the standard range with any score other than three on the raw scale due to their narrow range of responses. The first participant (left column), tends to give more varied answers, so their standard range is much wider, spanning from three to five. Their mean scores are also quite different, as the mean for first participant is between four and five while the second participant’s mean is approximately three.

Many EMA variables of interest, such as negative affect and attention, are notoriously skewed. Figure 3.2-A shows the skewed distribution of raw attention scores for all participants. After being standardized using participant-specific
means and variances, the distribution of standardized attention scores, shown in Figure 3.2-B, is much more symmetric. This allows us to avoid some of the typical issues that arise when dealing with significantly skewed distributions.

Among the various measures of within-participant stability, one of the most widely used is the MSSD (Jahng, Wood, and Trull, 2009), defined as

$$MSSD_i = \frac{1}{T_i - 1} \sum_{j=1}^{T_i-1} (x_{ij+1} - x_{ij})^2,$$

where $x_{ij}$ and $x_{ij+1}$ are the successive responses for participant $i$ at times $j$ and $j + 1$, and $T_i$ is the total number of observations for participant $i$. MSSD is directly related to the correlation between two successive observations (Jahng, Wood, and Trull, 2009). In our analyses, we calculate the MSSD for each participant using their normalized EMA scores and show that the fragmentation measure proposed here provides information about the within-participant stability beyond what is provided by the MSSD.

### 3.2.5 Fragmentation Measures

Fragmentation measures have previously been used in studies of accelerometry-measured physical activity to study patterns of accumulation of sedentary and active time by characterizing the duration of sedentary and active "bouts," or frequency of transitions, between sedentary and active states (Chastin and Granat, 2010; Di et al., 2017; S.P. Lim et al., 2011; Paraschiv-Ionescu, Buchser, and Aminian,
2013). These methods assess whether the total sedentary or active time is accumulated via a smaller proportion of longer bouts or a larger proportion of bouts of approximately equal durations. Fragmentation measures in this study were calculated via the fragmentation function from the actigraphy package in R (Di, Muschelli, and Zipunnikov, 2018; see example code for details). This framework can be readily applied to study any binary time series data.

Here, we have defined the two states as being inside or outside the participant-specific standard range. A participant is inside the standard range (ISR) when the response for the given EMA variable is less than one standard deviation from the participant-specific mean. Conversely, a participant is outside the standard range (OSR) when the response is more than one standard deviation from the mean. Thus, ISR and OSR correspond to assessments when normalized scores are respectively inside and outside of the $(-1, 1)$ interval.
Our fragmentation measure, denoted by $\lambda$, is the reciprocal of the average bout duration for each specific state. It can be shown that $\lambda$ is equal to the conditional probability of being in one state and transitioning to the other and as such, it will be a positive number between 0 and 1 (Di et al., 2017). Larger (and smaller) values of $\lambda$ will correspond to more (and less) frequent transitioning from the current state, respectively. Explicitly, if we define $y_t$ to be the current state and $y_{t+1}$ to be the state at the next assessment, $\lambda_s = P(Y_{t+1} = OSR | y_t = ISR)$, or the probability that the participant transitions to outside their standard range, given that they are currently inside their standard range (Di et al., 2017). Subscripts $s$ and $o$ denote whether the parameter refers to transitioning from inside-to-outside ($s$) or from outside-to-inside ($o$) the standard range. As shown in Di et al., 2017, these transition probabilities can be estimated as

$$
\lambda_o = \frac{n_o}{t_o},
$$

$$
\lambda_s = \frac{n_s}{t_s},
$$

(3.1)

where $n_s$ and $n_o$ are respectively the number of bouts spent in and outside the standard range, and $t_s$ and $t_o$ are the total time spent in each respective state. Assessments for each participant are taken four times per day at irregularly spaced intervals, the time $t_s$ is the number of assessments with a score in the participant-specific standard range, $t_0$ follows similarly. It is important to note that "time ($t$)" in this case is discrete and corresponds to the assessments done four times a day. Therefore, our method may potentially miss short-term fluctuations that can happen between the assessments. For example, if a participant is in the standard
range at two successive assessments, we would not be able to detect any out of range changes between the assessments. Thus, fragmentation metrics rely on the sampling frequency and timing of the assessments. Bias attributable to missingness could be a consequence of the fragmentation approach if the missingness is non-random between key study subgroups.

The definition of the two states is an important feature of this approach, but the fragmentation values are also dependent upon the definition of the standard range which is defined by the cutoff. In this application, we use the cutoff of one participant-specific standard deviation, but if the range were increased to two or three standard deviations, the conditional probability of transitioning out of the standard state would be lower and conversely, the conditional probability of transitioning into the standard state would be higher. This would change the distribution of values for $\lambda$.

The estimated parameter $\lambda$ often has a symmetric distribution in various samples and is more intuitive than alternative fragmentation measures (Chastin et al., 2011; Di et al., 2017). If both $\lambda$s are large, the trajectory is more fragmented, with shorter average bout durations. If both $\lambda$s are small, the trajectory is less fragmented (longer bout durations). Estimated probability density functions of the $\lambda$s both inside and outside the standard range, are shown in Figure 3.3 for both mood and attention. The color of each curve corresponds to the mood disorder subgroup. Inside the standard range, $\lambda$ tends to be smaller for both mood and attention (C and D), whereas for outside, the values tend to be larger (A and B). As $\lambda$ is the inverse of the average bout duration, the opposite is true for the average
bout durations of mood and attention. The average bout durations for each mood disorder subgroup, shown in Table 3.2, are larger ISR, typically around 4 to 6 assessments, than OSR, typically around one and one half and two assessments. They are similar for both mood and attention and tend to be longer and more variable for individuals in the control group when inside the standard range (ISR). To be explicit, an average "duration" for individuals with BP1 of 4.1 (see Table 3.2) means that individuals remain inside the standard range of mood for an average of 4.1 consecutive assessments. Outside the standard range of mood, the average duration for BP1 is only 1.7 assessments.

Table 3.2: Average Bout Duration for Mood and Attention Inside the Standard Range (ISR) and Outside the Standard Range (OSR) by Mood Disorder Subgroup

<table>
<thead>
<tr>
<th>Group</th>
<th>Mood ISR M (SD)</th>
<th>OSR M (SD)</th>
<th>Attention ISR M (SD)</th>
<th>OSR M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar I</td>
<td>4.1 (2.0)</td>
<td>1.7 (.5)</td>
<td>4.1 (3.5)</td>
<td>1.8 (.6)</td>
</tr>
<tr>
<td>Bipolar II</td>
<td>3.8 (2.1)</td>
<td>1.7 (.4)</td>
<td>3.8 (2.0)</td>
<td>1.6 (.9)</td>
</tr>
<tr>
<td>MDD</td>
<td>4.9 (3.6)</td>
<td>1.6 (.5)</td>
<td>4.4 (3.2)</td>
<td>1.6 (.5)</td>
</tr>
<tr>
<td>Control</td>
<td>5.8 (4.2)</td>
<td>1.6 (.5)</td>
<td>5.4 (4.0)</td>
<td>1.6 (.5)</td>
</tr>
</tbody>
</table>

Note: Average bout duration is defined as the average number of assessments that an individual remains within a specific state (ISR or OSR, in this case). Assessments are given four times per day. MDD = Major Depressive disorder.

To illustrate the potential that fragmentation measures have to provide information above and beyond that contained in the MSSD, we can look at four concrete examples of data from participants with specific MSSD values and varying fragmentation levels. In Figure 3.4, whereas both participants have similar MSSD around five to five and one half, the corresponding $\lambda_o$ values are quite different,
Figure (3.3) Fragmentation Metric Distributions: Distribution of calculated $\lambda$s by mood disorder subgroup. The upper and lower left panels are $\lambda_0$ and $\lambda_s$ for attention respectively, and the upper and lower right panels are the densities of $\lambda_0$ and $\lambda_s$ for mood.

with the participant on the left having $\lambda_0 = 0.18$ (a low probability of transitioning to ISR) and the participant on the right having $\lambda_0 = 0.92$ (a high probability of transition to ISR). Both participants have quite variable responses with wide ranges, but the participant on the left has long bouts or periods of time when they remain in the OSR state, whereas the second participant does not stay in the OSR, but transitions back to the standard range very frequently, within one or two assessments typically.
In Figure 3.5, we included two participants whose assessment scores hardly vary and have a small MSSD (around 0.7). Even though the MSSD values are similar, the $\lambda_s$ values are quite different. The first on the left has $\lambda_s = 0.82$ (a high likelihood of transitioning from the standard state), and the participant on the right has $\lambda_s = 0.06$ (a low likelihood of transitioning from the standard state). The binarized responses in the bottom plot for each participant show the participant on the right has only a few bouts in the standard range and they tend to be longer than the first participant who has many short bouts in the standard range state. These examples show the potential fragmentation measures have to highlight unique aspects of stability that are not captured by MSSD.

### 3.2.6 Missing Data

On average, 17.4% of the mood and attention scores from the momentary assessments are missing for each participant. The 25th and 75th percentiles of missingness were 5.7% and 24.1%, with a standard deviation of 15.2%. The average missingness for each participant’s EMA response by diagnostic group was not significantly different ($p = 0.72$): BP1 17.9% (SD = 19.4%), BPII 15.9% (SD = 13.4%), MDD 16.5% (SD = 14.1%), and Control 18.2% (SD = 15.4%).

In most cases, missing values for the mood or attention score means that the participant did not respond to that particular survey and we have no information about the participant at that time.

To account for this, missing observations were removed by concatenating the two sides of the binary series. For example, if a participant had three consecutive
Figure (3.4) Large MSSD Examples: These two participants have high mean squared successive difference (MSSD) but different fragmentation levels ($\lambda_\alpha$). The participant on the left is less likely to transition out of the OSR (outside the standard range) state whereas the second participant on the right tends to spend shorter amounts of time OSR and is more likely to transition out of that state. However, both participants’ raw scores are highly variable and the MSSD is approximately the same for both.

assessments where their response stayed inside the standard range, then missed two assessments, followed by two more assessments with a response in the standard range, the length of this bout would be five assessments. This assumes that there is no transition out of the current state when an assessment is missed. If the two assessments after the missing values were outside the standard range, the length of the bout inside the standard range would be three assessments. Thus, the missing assessments are not included in the calculation of fragmentation.
Figure (3.5) Small MSSD Examples: These two participants have low mean squared successive difference (MSSD) but different fragmentation levels. The first participant on the left has a high $\lambda_s$ and is more likely to transition out of the standard range than the second participant on the right, who has a low $\lambda_s$ and is more likely to stay in the standard range.

### 3.2.7 Data Modeling

Following the normalization and calculation of the fragmentation indices for each participant, we fit a linear model, using $\lambda_s$ and $\lambda_o$ as the response. Models were fit for both EMA variables (mood and attention). We define the model for ISR as follows:

$$\lambda_s = \beta_0 + \beta_1 \times \text{BPI} + \beta_2 \times \text{BPII} + \beta_3 \times \text{MDD} + \beta_4 \times \text{MSSD} + \beta_5 X + \epsilon \quad (3.2)$$
\( \beta_{5}X \) corresponds to the other demographic variables used as covariates, including age, sex, and BMI. \( \beta_{0} \) is the baseline corresponding to the control group. The model for the OSR time is defined similarly. Including MSSD in the model allows us to assess the utility of \( \lambda \) beyond information offered by this traditional measure of stability. We are primarily concerned with \( \beta_{1}, \beta_{2}, \) and \( \beta_{3}, \) the model terms that correspond to the difference in average duration within a state for each mood disorder group compared with the control group (\( \beta_{0} \)).

The models were fit in R (R Core Team, 2018). Fragmentation measures were calculated using the fragmentation() function as described in Section 3.2.5.

### 3.3 Results

Table 3.3 contains the summaries of the number of assessments, or "time," with responses in and outside the standard range of each EMA variable for controls and the three mood disorder subgroups. From this table, we see that the control group tends to be inside the standard range more than the BPI, BPII, and MDD groups for both variables, 73% to 74% compared with 64% to 69%.

Table 3.4 provides a summary of the significant results from model fitting (Equation 3.2). The BPI group is significantly more fragmented inside the standard range (shorter ISR bouts) of attention (\( p = 0.004 \)) and is significantly less fragmented (longer OSR bouts) outside the standard range of attention (\( p = 0.021 \)) than control group participants, indicating that when individuals with BPI are within standard range of attention, the bouts are shorter on average, or more fragmented, than
### Table 3.3: Summary of Time Spent Inside the Standard Range (ISR) and Outside the Standard Range (OSR) by EMA Variable and Mood Disorder Subgroups

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>Avg. number of assessments ISR - M (SD)</th>
<th>OSR - M (SD)</th>
<th>Avg. proportion of assessments ISR - M (SD)</th>
<th>OSR - M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention</td>
<td>Bipolar I</td>
<td>15.1 (9.1)</td>
<td>27.9 (12.0)</td>
<td>35.6%</td>
<td>64.4%</td>
</tr>
<tr>
<td></td>
<td>Bipolar II</td>
<td>13.3 (6.3)</td>
<td>28.1 (8.8)</td>
<td>32.1%</td>
<td>67.9%</td>
</tr>
<tr>
<td></td>
<td>MDD</td>
<td>13.4 (7.4)</td>
<td>30.1 (10.3)</td>
<td>31.1%</td>
<td>68.8%</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>11.4 (7.4)</td>
<td>31.1 (10.0)</td>
<td>26.8%</td>
<td>73.2%</td>
</tr>
<tr>
<td>Mood</td>
<td>Bipolar I</td>
<td>13.3 (7.3)</td>
<td>29.2 (9.9)</td>
<td>30.7%</td>
<td>69.3%</td>
</tr>
<tr>
<td></td>
<td>Bipolar II</td>
<td>13.3 (5.6)</td>
<td>28.0 (9.2)</td>
<td>32.7%</td>
<td>67.3%</td>
</tr>
<tr>
<td></td>
<td>MDD</td>
<td>12.7 (7.4)</td>
<td>30.8 (10.3)</td>
<td>29.3%</td>
<td>70.7%</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>11.0 (6.9)</td>
<td>31.5 (10.0)</td>
<td>26.2%</td>
<td>73.8%</td>
</tr>
</tbody>
</table>

Note: Values are based on four survey periods per day.
EMA = Ecological Momentary Assessments; MDD = Major Depressive Disorder.

those in the control group. However, the bouts outside the standard attention levels (being either very inattentive or very attentive) are longer than in controls. Alternatively, the $\beta$ coefficient of 0.09 from the first line in Table 3.4 indicates there are 0.09 more bouts per assessment for those with BPI than controls inside the standard range. This translates to an average of approximately five more bouts than controls over the 2-week assessment period. The other $\beta$ values can be interpreted similarly.

The MDD group had significantly shorter bouts than controls inside the standard range of attention and mood ($p = 0.017, 0.039$) indicating that bouts spent in the standard range for mood and attention are shorter and are less stable than those in the control group. No significant differences were found in the average duration in the mood domain for those with BPI or BPII, either inside or outside the standard range of attention. The only other significant effect is the MSSD for attention ISR. The coefficients and p values for all other models for both mood and attention models are shown in Table 3.5 and Table 3.6, respectively.
### Table 3.4: Significant Fragmentation Effects of Mood Disorder Subgroups for Attention and Mood Domains

<table>
<thead>
<tr>
<th>EMA variable</th>
<th>Level</th>
<th>$\beta^{[95%CI]}$</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bipolar I</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention</td>
<td>ISR</td>
<td>0.09[0.03, 0.15]$^b$</td>
<td>0.003</td>
</tr>
<tr>
<td>Attention</td>
<td>OSR</td>
<td>$-0.09[-0.17, -0.01]^c$</td>
<td>0.022</td>
</tr>
<tr>
<td><strong>MDD</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention</td>
<td>ISR</td>
<td>0.05[0.01, 0.09]</td>
<td>0.003</td>
</tr>
<tr>
<td>Mood</td>
<td>ISR</td>
<td>0.04[0.002, 0.081]</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Note: EMA = Ecological Momentary Assessments; MDD = Major Depressive Disorder. ISR = inside the standard range; OSR = outside the standard range.  
$^a$ Coefficients for mood disorder subgroups from the linear model with the fragmentation measure, $\lambda$, as the response. Positive $\beta$ coefficients indicate that those in the mood disorder subgroup are more fragmented or that the bouts of time spent either in or out of the standard range for attention or mood are shorter on average than for those in the control group.  
$^b$ This $\beta$ coefficient indicates that those with BPI have, on average, 0.09 more bouts per assessment (an electronic diary entry) inside the standard range of attention than those in the control group. Over the entire 2-week study (56 assessments), these participants would have approximately five more bouts inside the standard range of attention than control participants.  
$^c$ Indicates that those with BPI have, on average, 0.09 fewer bouts per assessment than those in the control group when outside the standard range of attention. Over the 2-week study, they would have approximately five fewer bouts with attention outside the standard range.

Figures 3.6 and 3.7 examine the MSSD, $\lambda_o$, and $\lambda_s$ for attention and mood by mood disorder subgroup (BPI is shown in red, BPII in orange, MDD in green, and controls in blue). In the top left plot of Figure 3.6 ($\lambda_o$ vs. $\lambda_s$), there is a slight negative association between $\lambda_o$ and $\lambda_s$ across all groups. However, in the top left plot of Figure 3.7, the negative association between $\lambda_o$ and $\lambda_s$ is not as salient as in
Table 3.5: Statistics for Mood Fragmentation Models

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient ($\beta$)</th>
<th>t</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.001</td>
<td>1.81</td>
<td>0.072</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>-0.014</td>
<td>-0.64</td>
<td>0.523</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.003</td>
<td>-1.91</td>
<td>0.057</td>
</tr>
<tr>
<td>BPI</td>
<td>0.008</td>
<td>0.22</td>
<td>0.825</td>
</tr>
<tr>
<td>BPII</td>
<td>-0.042</td>
<td>-1.27</td>
<td>0.205</td>
</tr>
<tr>
<td>MDD</td>
<td>0.005</td>
<td>0.21</td>
<td>0.832</td>
</tr>
<tr>
<td>Race (Black)</td>
<td>0.006</td>
<td>0.26</td>
<td>0.798</td>
</tr>
<tr>
<td>Race (Other)</td>
<td>-0.030</td>
<td>-0.66</td>
<td>0.513</td>
</tr>
<tr>
<td>MSSD</td>
<td>0.003</td>
<td>0.25</td>
<td>0.806</td>
</tr>
<tr>
<td>ISR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.0003</td>
<td>-0.67</td>
<td>0.504</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>0.007</td>
<td>0.38</td>
<td>0.702</td>
</tr>
<tr>
<td>BMI</td>
<td>0.001</td>
<td>0.56</td>
<td>0.575</td>
</tr>
<tr>
<td>BPI</td>
<td>0.033</td>
<td>0.99</td>
<td>0.325</td>
</tr>
<tr>
<td>BPII</td>
<td>0.047</td>
<td>1.60</td>
<td>0.111</td>
</tr>
<tr>
<td>MDD</td>
<td>0.042</td>
<td>2.04</td>
<td>0.042*</td>
</tr>
<tr>
<td>Race (Black)</td>
<td>-0.004</td>
<td>-0.17</td>
<td>0.865</td>
</tr>
<tr>
<td>Race (Other)</td>
<td>0.027</td>
<td>0.68</td>
<td>0.497</td>
</tr>
<tr>
<td>MSSD</td>
<td>-0.001</td>
<td>-0.08</td>
<td>0.935</td>
</tr>
</tbody>
</table>

Note: BMI = body mass index; BPI = Bipolar I; BPII = Bipolar II; MDD = Major Depressive Disorder. ISR = inside the standard range; OSR = outside the standard range; MSSD = mean squared successive difference.

* $p < 0.05$

Figure 3.6. In the top right plots of Figures 3.6 and 3.7, there is no evidence of a significant relationship between $\lambda_0$, MSSD, and the mood disorder subgroups.

In the lower plot of Figure 3.6, there are very few individuals in the BPI group that have high MSSD values, and none have low $\lambda_s$ values. Conversely, for the mood responses in the lower-left plot in Figure 3.7, there are only one or two
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient (β)</th>
<th>t</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.001</td>
<td>-1.52</td>
<td>0.129</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>-0.010</td>
<td>-0.42</td>
<td>0.673</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.0001</td>
<td>0.04</td>
<td>0.966</td>
</tr>
<tr>
<td>BPI</td>
<td>-0.092</td>
<td>-2.31</td>
<td>0.022*</td>
</tr>
<tr>
<td>BPII</td>
<td>-0.059</td>
<td>-1.63</td>
<td>0.104</td>
</tr>
<tr>
<td>MDD</td>
<td>-0.030</td>
<td>-1.19</td>
<td>0.235</td>
</tr>
<tr>
<td>Race (Black)</td>
<td>0.024</td>
<td>0.98</td>
<td>0.330</td>
</tr>
<tr>
<td>Race (Other)</td>
<td>-0.038</td>
<td>-0.83</td>
<td>0.409</td>
</tr>
<tr>
<td>MSSD</td>
<td>-0.005</td>
<td>-0.69</td>
<td>0.492</td>
</tr>
<tr>
<td>ISR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.001</td>
<td>-1.16</td>
<td>0.248</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>-0.007</td>
<td>-0.37</td>
<td>0.708</td>
</tr>
<tr>
<td>BMI</td>
<td>0.0002</td>
<td>0.17</td>
<td>0.867</td>
</tr>
<tr>
<td>BPI</td>
<td>0.094</td>
<td>2.96</td>
<td>0.003**</td>
</tr>
<tr>
<td>BPII</td>
<td>0.032</td>
<td>1.13</td>
<td>0.261</td>
</tr>
<tr>
<td>MDD</td>
<td>0.048</td>
<td>2.45</td>
<td>0.015*</td>
</tr>
<tr>
<td>Race (Black)</td>
<td>0.009</td>
<td>0.43</td>
<td>0.664</td>
</tr>
<tr>
<td>Race (Other)</td>
<td>-0.032</td>
<td>-0.89</td>
<td>0.376</td>
</tr>
<tr>
<td>MSSD</td>
<td>0.022</td>
<td>4.03</td>
<td>0.0001***</td>
</tr>
</tbody>
</table>

Note: BMI = body mass index; BPI = Bipolar I; BPII = Bipolar II; MDD = Major Depressive Disorder. ISR = inside the standard range; OSR = outside the standard range; MSSD = mean squared successive difference.

* p < 0.05; ** p < 0.005; *** p < 0.005

observations in the BPI group with a small λ_s and a large MSSD. In particular, there is only one individual with a large λ_s among those with BPI disorder, regardless of the MSSD. Clearly, individuals in the BPI group tend to have small λ_s and small MSSD values, a tendency which is not apparent among the controls and other mood disorder subgroups.
Figure (3.6) Attention Fragmentation and MSSD: Comparison of outside the standard range (OSR) and inside standard range (ISR) fragmentation measures (top left). Comparison of OSR fragmentation measure and mean squared successive difference (MSSD; top right). Comparison of ISR fragmentation measure and MSSD (bottom left). The correlation between the variables on the x- and y-axes is shown above each plot on the top line. The slopes and respective p-values of the fitted lines for each mood disorder subgroup are on the second line above each plot.
Figure (3.7) Mood Fragmentation and MSSD: Comparison of outside the standard range (OSR) and inside standard range (ISR) fragmentation measures (top left). Comparison of OSR fragmentation measure and mean squared successive difference (MSSD; top right). Comparison of ISR fragmentation measure and MSSD (bottom left). The correlation between the variables on the $x$— and $y$-axis is shown above each plot on the top line. The slopes and respective p-values of the fitted lines for each mood disorder subgroup are on the second line above each plot.
3.4 Discussion

This work proposes fragmentation measures of stability for EMA data and demonstrates that these novel measures provide information beyond what can be accounted for by standard participant-specific measures such as mean, variance, and MSSD. Using this novel approach to characterize stability of mood and attention levels, we showed that the stability of the EMA differs among those with BPI and MDD compared with controls. Specifically, those with BPI exhibit less stability than controls inside the standard range of attention, but greater stability and more assessment responses outside the standard range of attention. Those with MDD also exhibited less stability than controls inside the standard range of both mood and attention.

Concerning clinical relevance, the finding that people with BPI tend to have greater fragmentation in attention with frequent ratings of distractibility suggests that the fluctuation of depressed mood is not a salient feature of this condition. This replicates clinical studies that highlight that attention is a more important psychological process in BPD than mood (Scott et al., 2016; Johnson, Gershon, and Starov, 2015). Although the instability of mood in people with a history of MDD may be expected, these findings demonstrate patterns of fluctuation in real time that can be used to define targets for treatment designed to stabilize mood. These findings also inform classification of mood disorders by distinguishing between emotional regulatory patterns among those with BPD versus those with MDD and controls. The present analyses focused on $\lambda$ since it is one of the most
informative and easily interpretable fragmentation measures. However, other fragmentation measures are available such as the Gini index and parameters of power law distribution (Di et al., 2017). Additionally, it should be noted that the distribution of fragmentation values is sensitive to the definition of the two states and other cutoffs such as above/below -2, -1, 1, and 2 standard deviations could be employed.

The findings should be interpreted relative to the specific characteristics and limitations this study. One limitation of the data is the moderate number of BPI participants \((n = 35)\) compared with the controls \((n = 159)\). Additionally, as each participant could generate a maximum of 56 observations (four per day over two weeks), individuals with low variability of mood or attention levels have very few observations outside of the standard range, increasing the uncertainty for estimates of their fragmentation. Conversely, if individuals with high variability have frequent short-term fluctuations between assessments, we may potentially miss changes between states due to the fact that individuals are surveyed only four times per day. Thus, fragmentation metrics are sensitive to the sampling frequency and timing of the assessments.

The validity of the results obtained in this sample should be tested over longer periods of assessment, as well as with additional types of mental disorders. A final consideration is the inability of these fragmentation measures to determine whether differences in stability between mood disorder subgroups are related to high or low EMA variable scores (i.e., if a subgroup is more stable when their mood and attention levels are either high or low). The definition of future work
should involve the extension of the proposed fragmentation measures to account for directionality in the EMA responses across the subgroups.

This work demonstrates that fragmentation methods can characterize stability of specific psychological states beyond that of typical measures of variability and stability. The novel substantive finding that fluctuations in attention may be a more salient core domain than mood fluctuations in people with BPI disorder also highlights the utility of the fragmentation approach in examining variability of state measures in psychiatric disorders, as well as other chronic diseases.
References


Chapter 4

Variable Domain Functional Principal Component Analysis

4.1 Introduction

We propose a functional principal component technique for variable-domain functional data. Variable-domain functional data have become increasingly common in health studies. For example, during inpatient hospital stays, subject measurements, such as disease severity or blood pressure, are collected regularly (Gellar et al., 2014). The length of the hospital stay could vary from a day to weeks or months depending on the medical condition and potential complications. Other examples include EEG-power spectrum in adjacent 30-second intervals during sleep (Crainiceanu et al., 2009) and acceleration profiles during chair stands (Bai et al., 2012; He et al., 2014). Indeed, the duration of sleep and of standing-up from a chair varies both within and especially between subjects and naturally leads to functional data with a variable domain.
To be concrete, we focus here on the Improving Care of Acute Lung Injury Patients (ICAP) study (Needham et al., 2006). Information was collected for each of the 520 subjects at enrollment, once per day while staying in the ICU, up until death or discharge. Out of the 520 patients, 283 (54%) survived in the ICU and were discharged.

One measure of particular interest was the Sequential Organ Failure Assessment (SOFA) score. SOFA is a sum of scores from six physiological system assessments (respiratory, coagulation, liver, cardiovascular, renal, and central nervous system). For each physiological system, assessments were given once daily and the scores varied from 0-4 where larger values indicate worse performance of the system. Thus, SOFA scores vary between 0 and 24 with larger values indicating worse overall health status.

To provide a better understanding of the data, Figure 4.1-A displays the SOFA scores for subjects separated by group (those who died and those who survived and were discharged); for presentation purposes, data are not shown beyond 100 days. Trajectories are ordered by the length of stay for each subject. Figure 4.1-B displays the trajectories for all subjects in each group (gray solid lines), while five subjects are highlighted as black solid lines. Figure 4.1-C displays the distribution of the length of stay for each group on the log scale.

The functional domain of the SOFA score varies from one day to more than 150 days. Given this data structure, we are interested in finding the main patterns of variation in SOFA scores as a function of time after being admitted to the ICU. Principal component analysis (PCA) is a common approach for identifying the
Figure (4.1) SOFA scores for individuals with acute lung injury (ALI) in the ICU:  

A: Patients are divided into two groups, those that died (bottom) and those that survived and were discharged (top). Each row in the plot depicts the SOFA scores measured once daily for each subject and subjects are ordered by the length of time in the hospital, darker red indicates higher SOFA scores. B: SOFA score trajectories from five subjects are highlighted in black. Scores for all other subjects are shown in gray. C: Estimated empirical distribution of length of stay on log scale for all subjects; patients who survived and were discharged are represented by the distribution in red and patients who died are represented by the distribution in green.

...main directions of variation in functional data; however, the uneven domain of the function (patient-specific length of stay) makes a brute-force approach impossible. One possible alternative would transform the data to the same domain, interpolate the data on a common grid, and conduct PCA on the transformed data. While this
approach is simple, in many situations it is scientifically invalid. For example, in the ICU data from the ICAP study it would make little sense to register the data for a subject who dies after two days in the ICU to the data of a subject who survived and was discharged after 60 days in the ICU.

These concerns are not limited to the ICU data. In sleep studies, the electroencephalograph records electric activity in the brain for the duration of sleep, which can vary substantially by subject. Registering each subject’s sleep cycle to a common domain could adversely distort the observed patterns of sleep, including the duration of RK sleep stages, such as rapid eye movement (REM) or non-REM sub-stages.

Another possible alternative would be to use “conditional fPCA” and condition on the domain length after transforming the data to the same domain (Cardot, 2007). This approach would allow the principal components to vary according to domain lengths while ameliorating some concerns of scientific validity when the data are registered to the same domain. However, the smoothed component estimates can yield, theoretically and in practice, widely varying results for the transformed data compared to estimates on the original scale. This is due to the difference in penalty for the transformed versus untransformed true principal components when the transformation is not a simple linear rescaling (see the Appendix for details). To address these issues, we propose a method inspired by PCA but tailored to functional data with varying domains.

A simple method for estimating the principal sources of variation conditional on the length of the domain would be to separate the functional domain into
strata according to the length of the domain. This would group curves of similar length, transform them onto the same scale, and analyze each stratum separately. However, selecting the number and size of the strata is a difficult task and ignores information across strata. In order to account for this, we propose to estimate a 3-dimensional (3D) covariance operator that accounts for the length of the domain, using a 3D penalized thin-plate spline smoother to borrow strength across domain lengths.

The remainder of the paper is organized as follows. Section 4.2 proposes FPCA to data with variable domain and non-parametric estimation methods for the conditional mean and variance. Section 4.3 provides a detailed simulation study to evaluate the proposed method. Section 4.4 describes the application of the proposed methods to two real examples. Section 4.5 provides the discussion and conclusions.

4.2 Methods

4.2.1 Variable-Domain Approach

We propose variable-domain functional principal component analysis (vd-FPCA). This approach is related to conditional FPCA (Cardot, 2007). However, here we focus on the unique case where the conditioning variable is the domain length.

Let \((Y, M)\) be random variables where \(M \in \mathcal{T}\) is the domain length and \(Y \in H = L^2(\mathcal{T})\) is the random function on the compact interval \(\mathcal{T} = [0, T]\). In the
case that $\mathcal{T}$ is not of the form $[0, T]$, this method will still be applicable; however, we specify this form for the interval for the sake of simplicity. We define the expectation $\mu(t, m)$ and covariance operator $\Gamma^m$, conditional on $M = m$ as

$$
\mu(t, m) = E[Y(t)|M = m]
$$

$$
\Gamma^m f(t) = \int \gamma(t, s, m)f(s)ds, \quad t, s \in [0, m], \quad m \in \mathcal{T}.
$$

Here,

$$
\gamma(t, s, m) = \text{cov}[Y(t), Y(s)|M = m]
$$

$$
= E\{[Y(t) - \mu(t, M)]\{Y(s) - \mu(s, M)\}|M = m\}, \quad t, s \in [0, m],
$$

are the entries of the 3D covariance operator, where $0 \leq t, s \leq m$ and $m$ is the length of the domain. The conditional principal component and eigenvalue $\psi_k(t, m)$ and $\lambda_k(m)$ are the kth orthonormal eigenfunction and eigenvalue of $\Gamma^m$, respectively. The covariance function $\gamma(t, s, m)$ can be expressed as:

$$
\gamma(t, s, m) = \sum_k \lambda_k(m)\psi_k(s, m)\psi_k(t, m), \quad t, s \in [0, m].
$$

The eigenfunctions $\psi_k(t, m)$ are orthonormal for a given $m$, that is $\int_0^m \psi_k(t, m)\psi_{k'}(t, m)dt = 1$ when $k = k'$, and 0 otherwise. However, the eigenfunctions are not required to be orthonormal across $m \in [0, T]$.

Using the conditional Karhunen-Loéve expansion, we can obtain the best linear approximation $\tilde{Y}^K(t, m)$ for the function $Y(t, m)$ of finite dimension $K$ as:

$$
\tilde{Y}^K(t, m) = \mu(t, m) + \sum_{k=1}^K \xi_k(m)\psi_k(t, m), \quad (4.1)
$$

94
where $\xi_k(m)$ is the $k^{th}$ principal component score of $Y$ and $\psi_k(t, m)$ is the $k^{th}$ orthonormal eigenfunction of the covariance operator $\Gamma^m$, for $k = 1, \ldots, K$ (Loeve, 1978; Cardot, 2007).

4.2.2 Estimation

We can estimate both the conditional expectation $\mu(t, m)$ and the conditional covariance function $\gamma(t, s, m)$ using non-parametric smoothing methods. Cardot (2007) proposed using kernel smoothers to estimate these functions; however, the method of Cardot does not have readily available software and has not been used extensively in the literature because of the manual tuning of parameters and its complexity. We use penalized thin plate splines (PTPS) and the function `gam` from the `mgcv` package in R, which has been extensively developed to fit a variety of spline models and utilizes fully automated methods to select the amount of smoothing (Wood, 2003; Wood, 2006; Wood, 2018; R Core Team, 2018).

PTPS are particularly useful in this case because they can be fit over multiple variables and do not require the surface to be rectangular. Moreover, PTPS do not require manually choosing knots, basis functions, or manipulation of the smoothing parameter; all parameters can be estimated directly (Wood, 2006). PTPS are isotropic, meaning that the basis functions are symmetric in all directions. Though this could potentially cause problems if the optimal smoothness is different in the two directions ($t$ and $M$), we do not anticipate noticeable problems in practice since the range in both directions is identical ($\max\{t\} = \max\{M\}$) and they both have related interpretations (Gellar et al., 2014).
We want to smooth the mean over the two-dimensional domain defined by \((t, m) : 0 \leq t \leq m\), which is a triangle, and the covariance operator over the three-dimensional domain defined by \((t, s, m) : 0 \leq t, s \leq m\), which is a rectangular pyramid. To better understand the differences, Figure 4.2 shows the distinction between the two cases. The two top panels display the shape of the conditional mean surfaces. The top-left panel shows the traditional case where the conditional mean is estimated over a rectangular domain. The top-right panel displays the variable-domain case, indicating that the surface is limited by the length of the domain, defining a triangular region.

The two lower panels in Figure 4.2 display the shape of the covariance function. The traditional case, functional estimation over a rectangular prism-like region, is on the left and the variable-domain case is on the right. We use PTPS to non-parametrically estimate both the mean and covariance functions (Wood, 2003).

To estimate the conditional mean \(\mu(t, m)\), we use the following bivariate smoothing model,

\[ Y_i(t) = \mu(t, m) + \epsilon_i(t), \]

where \(\epsilon_i(t) \sim N(0, \sigma^2)\) is the random error. Penalized thin plate splines find the function \(\hat{f}\) which minimizes

\[ ||Y - f||^2 + \lambda J_{md}(f), \tag{4.2} \]

where \(\lambda\) is the smoothing parameter, balancing the fit to the data and smoothness of the function, and \(J_{md}\) is a penalty function measuring the “wiggliness” of \(f\), (Wood, 2006). For simplicity, we do not include any additional covariates in the model,
Figure (4.2) Mean and Covariance Surface Depictions: Comparison of the mean and covariance surfaces in the variable domain (VD) case in the right column and the same domain (SD) case in the left column. The two figures in the top row are representations of conditional mean surfaces. Likewise, the two figures in the bottom row are the corresponding representations of the covariance function space, with the VD case on the right.

but in general, they can be included in the estimation of these functions. The smoothing parameter $\lambda$ can be estimated by generalized cross-validation (GCV) or restricted maximum likelihood (REML). We will use REML because it can be easily integrated in the mixed effects modeling and is readily generalizable to more complex models (Ruppert, Wand, and Carroll, 2003; Wood, 2011). We can perform this estimation, fitting the model in (4.2), using a generalized additive model fitting function, gam from the mgcv package in R (Wood, 2018; R Core Team, 2018).

We can then use a similar procedure to estimate the conditional covariance function, based on the following additive model

$$\{Y_i(t) - \mu(t, m)\}\{Y_i(s) - \mu(s, m)\} = \gamma(t, s, m) + e_i(t)$$  \hspace{1cm} (4.3)
where \( t, s \in [0, m] \) and \( e_i \) is the random error. We again use penalized thin plate splines to fit the model in (4.3). After estimating the conditional covariance function \( \hat{\gamma}(t, s, m) \) over \( M \), we can extract the covariance matrix for any specific \( m \) and diagonalize it using standard eigen-decomposition methods. The resulting eigen-function estimates, \( \hat{\psi}_k(t, m) \), are smooth over \( t \), conditioned on each \( m \) because the 3D covariance estimate, \( \hat{\gamma}(t, s, m) \), is smooth over \( t, s, \) and \( m \).

Returning to Equation (4.1), we know that the best linear representation \( \hat{Y}_K \) of \( Y \), given \( M \) in \( K \) dimensional space is,

\[
\hat{Y}_K(t, m) = \mu(t, m) + \sum_{k=1}^{K} \langle Y(t, m) - \mu(t, m), \psi_k(t, m) \rangle \psi_k(t, m).
\]

where \( \langle Y(t, m) - \mu(t, m), \psi_k(t, m) \rangle \) are the principal component scores \( \xi_k \) for \( i = 1, \ldots, K \), which have a mean of 0 and variance of \( \lambda_k \). It follows that we can estimate the scores by setting \( \hat{\xi}_k = \int_{0}^{m} (Y(t, m) - \hat{\mu}(t, m)) \hat{\psi}_k(t, m) dt \) and numerically integrating over the domain for each subject (Cardot, 2007).

### 4.2.3 Alternative Approaches

We compare the variable domain approach to two potential alternative methods. The first, or “scaled” method, takes each curve of varying length and stretches or shrinks it to the same time scale, interpolating linearly between the observed points so the functions are observed on a common grid or matrix. A typical functional PCA can then be performed on the resulting matrix (Ramsay and Silverman, 2005; Di et al., 2009; Goldsmith, Greven, and Crainiceanu, 2013). This method may have more scientific justification in cases where each observed function has similar
landmarks or features that occur over different time domains. Examples could include growth curves or accelerometer measurements of a subject walking, or standing from a chair.

The second alternative method we propose, the “weighted” method, can be viewed as having a data generating mechanism where the entire functional domain is censored, or only partially observed. This could occur when subjects are repeatedly measured during a follow up period but drop out early from the study, or if they have many missing visits. More precisely, a naive covariance matrix is first calculated by summing the crossproduct of the de-meaned data in each corresponding element of the covariance matrix. For example, if the maximum number of observations of all subjects is 100, the data for a subject with non-missing values for just the first three time points only contributes to the first $3 \times 3$ sub-matrix of the $100 \times 100$ covariance matrix (Staniswalis and Lee, 1998; Yao, Mueller, and Wang, 2005). The covariance matrix is then smoothed using penalized splines, weighting by the number of observations that contribute to each element of the matrix. We compare the variable-domain method to the two alternative approaches first using simulated data in Section 4.3, and then with two contrasting study examples in Section 4.4.
4.3 Simulation Study

4.3.1 Simulation Design

We evaluate the proposed variable-domain principal components analysis method first using simulated data. The generated data \((Y_i, m_i)\) consist of \(i = 1, \ldots, n\) independent and identically distributed (i.i.d.) realizations of the pair of random variables \(M\) (scalar) and \(Y\) (function). In this simulation, we consider the case with only one functional component without additional covariates. We specify the data generating model for subject \(i\) at time \(t_j\), where \(t_j \in [0, m_i]\) as

\[
Y_i(t_j, m_i) = \mu(t_j, m_i) + \sum_{k=1}^{K} \xi_{ik} \psi_k(t_j, m_i) + \epsilon_i(t_j),
\]

where \(\xi_{ik} \sim N(0, \lambda_k)\), \(\epsilon_i(t_j) \sim N(0, \sigma^2)\), and \(K = 10\). We sample \(t\) at a rate of 1 observation per unit time. Simulations are designed to compare the three approaches when the domain length is variable. We generate data and compare the performance of each method using all possible combinations of the following factors:

1. Two sample sizes \(N\): 100, 500.

2. Two distributions for \(M\): D1 - Uniform (2, 80), and D2 - Geometric\((p = 0.06)\).

3. Three levels of noise, \(\sigma\): low \((\sigma = 0.01)\), medium \((\sigma = 0.1)\), and high \((\sigma = 1)\).

4. Two eigenfunction options: A - functions linearly scale over domain length, B - a weighted combination of two eigenfunctions where weights are a function of domain length \(M\).
The eigenvalues are defined as \( \lambda_k = 0.5^{k-1} \) for \( k = 1, \ldots, 10 \) and the mean function as \( \mu(t, m_i) = 0.0001(t - 120)^2 + 3 \cdot \sin(\pi t / 60) \).

We evaluate the performance of the variable domain approach both in situations where the other approaches may be more natural (e.g. eigenfunctions are linearly scaled according to domain length \( M \) - Method A), as well as in situations where the variable domain approach may be more intuitive (e.g. shape of eigenfunctions varies non-linearly with respect to the domain length \( M \) - Method B). In Method A, eigenfunctions are linearly scaled over the domain length and are defined as,

\[
\psi_k^{(A)}(t, m_i) = \left\{ \frac{\sqrt{2}}{\sqrt{m_i}} \sin \left( \frac{2\pi t}{m_i} \right), \frac{\sqrt{2}}{\sqrt{m_i}} \cos \left( \frac{2\pi t}{m_i} \right), \frac{\sqrt{2}}{\sqrt{m_i}} \sin \left( \frac{4\pi t}{m_i} \right), \frac{\sqrt{2}}{\sqrt{m_i}} \cos \left( \frac{4\pi t}{m_i} \right), \cdots \right\}.
\]

The \( t/m_i \) and \( \sqrt{2}/\sqrt{m_i} \) terms stretch the time for each function and ensure that \( t/m_i \in [0, 1] \) for all \( m_i \). These functions are mutually orthonormal, conditional on \( m_i \). The first two eigenfunctions for \( m = 10, 30 \) and 60 are shown in Figure 4.3-A and 4.3-B. As the length of the domain increases, each eigenfunction stretches over the domain and decreases in amplitude.

By comparison, the eigenfunctions for Method B are transformed as the domain length \( M \) increases using a weighted combination of two different eigenfunctions. The eigenfunctions from Method B are defined as,
Figure (4.3) Eigenfunction Transformations: A: First eigenfunction ($k = 1$) linearly stretched (Method A) over domain lengths $m = 10, 30, \text{and } 60$. B: Second eigenfunction ($k = 2$) linearly stretched (Method A). C: First eigenfunction ($k = 1$) using weighted combination of two eigenfunctions (non-linear, Method B). In C, the eigenfunctions transform from the second eigenfunction in B at $m = 10$ to the first eigenfunction in A at $m = 60$.

\[
\begin{align*}
\psi_1^{(B)}(t, m_i) &= W(m_i) \frac{\sqrt{2}}{\sqrt{m_i}} \sin(2\pi t / m_i) + \{1 - W(m_i)\} \frac{\sqrt{2}}{\sqrt{m_i}} \cos(2\pi t / m_i) \\
\psi_2^{(B)}(t, m_i) &= W(m_i) \frac{\sqrt{2}}{\sqrt{m_i}} \sin(4\pi t / m_i) + \{1 - W(m_i)\} \frac{\sqrt{2}}{\sqrt{m_i}} \cos(4\pi t / m_i) \\
\psi_3^{(B)}(t, m_i) &= \cdots
\end{align*}
\]  

(4.5)

where the weighting function $W(m_i) = \Phi(m_i)$ is the normal CDF with $\mu_W = 30$ and standard deviation $\sigma_W = 10$. The weighting function is shown in Figure 4.4-B. It is overlaid on top of the two possible distributions for $M$ with the scale for the distribution on the right side axis. Distribution 1 (D1), shown in red, is Uniform(2,
80) and distribution 2 (D2), shown in gray, is Geometric(0.06). The density scale for the histograms is on the left side axis.

Figure (4.4) Simulation Setup: A: Trajectories for 500 simulated subjects. Four individual subjects are highlighted in various colors. Mean function is shown with dashed black line. B: Overlay of the two distributions of $M$ (D1 and D2) and the weighting function $W(m)$. D1, shown in red, is a Uniform(2, 80). D2, shown in gray, is Geometric(0.06).

To simulate observations for each subject $i$, we first generate $\zeta_i = \{\zeta_{i1}, \zeta_{i2}, \ldots, \zeta_{i10}\}$, $m_i$, and $\epsilon_i(t_j) = \{\epsilon_{i1}, \ldots, \epsilon_{ij}\}$ where $t_j = 0, 1, 2, \ldots, t_J = m_i$ using each of the previously specified distributions. Then, using $\zeta_i$, $m_i$, and $\epsilon_i$, we can calculate $X_i = \mu + \zeta_i' \Psi_i + \epsilon_i$ where $\mu$ is the mean vector and $\Psi_i$ is a $J \times 10$ matrix with each column being the corresponding true eigenfunction evaluated at $t_j$. Using this data generating mechanism, we simulate the trajectories of responses for each scenario. One sample of 500 curves simulated using $M \sim \text{Geometric}(0.06)$ and eigenfunctions that were linearly scaled is shown in Figure 4.4-A. Four individual
subject trajectories are highlighted in color while the remaining trajectories for all other subjects are in gray. The mean is represented by a dotted black curve.

4.3.2 Simulation Results

The three FPCA methods were compared in 16 distinct scenarios, using all possible combinations of the four factors detailed in Section 4.3.1. The approaches were compared using the average root mean squared error (ARMSE) for the data and for the principal components using 500 simulations. The two measures of ARMSE were calculated as follows:

\[
\text{ARMSE}_Y = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{1}{J} \sum_{j=1}^{J} \left( Y_i(t_j, m_i) - \hat{Y}_i(t_j, m_i) \right)^2}
\]  \hspace{1cm} (4.6)

\[
\text{ARMSE}_{PC}^{(k)} = \frac{1}{D} \sum_{d=1}^{D} \sqrt{\frac{1}{J} \sum_{j=1}^{J} \left( \psi_k(t_j, m_d) - \hat{\psi}_k(t_j, m_d) \right)^2}
\]  \hspace{1cm} (4.7)

where \( \hat{Y}_i(t_j, m_i) = \hat{\mu}(t_j, m_i) + \sum_{k=1}^{10} \hat{\xi}_{ik} \hat{\psi}_k(t_j, m_i) \) and the estimated components \( \hat{\psi}_k(t_j, m_i), \hat{\xi}_{ik} \) and \( \hat{\mu}(t_j, m_i) \) are calculated as detailed in Section 4.2.2. The ARMSE_{PC}^{(k)} is the average error relating to the \( k^{th} \) principal component for domain lengths \( m_d \in \{5, 10, 20, 30, 40, 50, 60, 70, 80\} \) and \( D \) is the number of domain lengths \( m_d \) used to calculate ARMSE_{PC}.

The ARMSE_Y values calculated using the simulated data are shown for each method and scenario in Table 4.1. Overall, the scaled and VD approaches have a much smaller ARMSE than the weighted approach. In addition, the distribution of
the domain length $M$ seems to have a large effect on the relative performance of each method. For example, at the lowest $\sigma$ level ($\sigma = 0.01$), the scaled approach does a better job of recapturing the data when the distribution of $M$ is uniform ($\text{ARMSE}_Y = 0.16$) compared to the VD method ($\text{ARMSE}_Y = 0.24$), especially when the eigenfunctions are scaled linearly with the domain length. When, the eigenfunctions are transformed non-linearly, the scaled approach still does better, but the difference is not as large as before (0.24 compared to 0.28). However, for skewed distributions of $M$, the VD approach has a smaller ARMSE (0.21, $N = 100$) than the scaled approach (0.28), particularly when the eigenfunctions depend non-linearly on the domain length.

As $\sigma$ is increased to 0.1, the scaled method performs about the same compared to the VD approach in the scenario when $M \sim \text{Uniform}$. Specifically, when we stretch the eigenfunctions linearly and $M$ has a uniform distribution, the ARMSE for the scaled approach and VD approaches is almost identical (0.31) when $n = 100$. When the eigenfunctions depend non-linearly on $M$, the ARMSE increases for both methods, but the VD method now has a slightly lower ARMSE (0.325) compared to the scaled approach (0.333) when $n = 100$. As was the case when $\sigma = 0.01$, the VD approach still has the smallest ARMSE of the three methods when $M$ has a skewed (Geometric) distribution and $\sigma = 0.1$.

At the highest level of noise ($\sigma = 1.0$), the ARMSE increases substantially for each of the approaches; however, the ARMSE for the VD approach has less of an increase. In fact, at high levels of $\sigma$, the VD approach has the lowest ARMSE of the three approaches by a wide margin.
Table 4.1: Average root mean squared error (ARMSE) from 500 simulations.

<table>
<thead>
<tr>
<th>Sample size (N)</th>
<th>100</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Geo</td>
<td>Uniform</td>
</tr>
<tr>
<td>Distribution (M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geo</td>
<td>0.201</td>
<td>0.213</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.210</td>
<td>0.278</td>
</tr>
<tr>
<td>Eigenfunction ~ M</td>
<td>Linear</td>
<td>NL</td>
</tr>
<tr>
<td>VD</td>
<td>0.226</td>
<td>0.233</td>
</tr>
<tr>
<td>Scaled</td>
<td>0.287</td>
<td>0.319</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.552</td>
<td>0.577</td>
</tr>
<tr>
<td>Method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VD</td>
<td>0.498</td>
<td>0.501</td>
</tr>
<tr>
<td>Scaled</td>
<td>0.767</td>
<td>0.768</td>
</tr>
<tr>
<td>Weighted</td>
<td>1.021</td>
<td>1.034</td>
</tr>
</tbody>
</table>

1 The distribution of domain lengths M is either Geometric(0.06) or Uniform(2, 80)
2 This is in reference to how eigenfunctions change as M increases. “Linear” refers to the eigenfunctions being linearly stretched over the domain length M. “NL” refers to the eigenfunctions being a weighted combination of two distinct eigenfunctions where the weight is a function of M. Details are found in Section 4.3.1

The ARMSE$_{PC}$ values comparing the estimated and actual principal components are shown in Table 4.2 for each scenario and each approach. The VD and the weighted methods have lower ARMSE than the scaled method in all scenarios. For the first PC, the ARMSE is slightly smaller for the weighted method (0.177) than the VD approach (0.183) for the cases when M ~ Geometric, $\sigma = 0.01$, N = 500, and the eigenfunctions are a non-linear function of M. However, when M ~ Uniform, the ARMSE is slightly lower for the VD method (0.165) than the weighted approach (0.182). For PC 2, the weighted approach does slightly better in most scenarios, but performs similar to the VD approach when M ~ Uniform and the eigenfunctions are a non-linear function of M. Results for the remaining PCs are analogous and are not shown.
An important component of the larger ARMSE for the scaled PC seems to be the magnitude of the function. Indeed, upon inspection of various iterations of a variety of simulation scenarios, the shape for the smaller domain lengths appears to provide a reasonable match to the true PC, at least in terms of visual inspection. However, the scale of the function is quite different from the true value of the PC and the difference gets worse as the domain length increases (the magnitude of the scaled PC remains the same for all domain lengths - while the magnitude of the true PC decreases). This difference in scale can be addressed when the corresponding scores are combined with the estimated PCs to reconstruct the data. This is likely a large part of the reason for the scaled approach performing well in terms of RMSE for data reconstruction, but not for the estimation of the actual component.

In summary, the VD method reconstructs the data better when the eigenfunctions are not linearly scaled across domain lengths or when the distribution of domain lengths is skewed. The performance of the the VD method is also less affected by an increase in measurement error. In addition, the VD method has a much lower ARMSE for reconstructing the principal components than the scaled approach across all scenarios and it performs either the same or very close to the weighted approach.


Table 4.2: Average root mean squared error (ARMSE$_{PC}$) for the first 2 principal components from 500 simulations in each scenario.

<table>
<thead>
<tr>
<th>Sample size (N)</th>
<th>100</th>
<th></th>
<th>500</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution (M)</td>
<td>Geo</td>
<td>Uniform</td>
<td>Geo</td>
<td>Uniform</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenfunction $\sim M$</td>
<td>Linear</td>
<td>NL</td>
<td>Linear</td>
<td>NL</td>
<td>Linear</td>
<td>NL</td>
</tr>
<tr>
<td>Method</td>
<td>PC 1, $\sigma = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VD</td>
<td>0.210</td>
<td>0.192</td>
<td></td>
<td>0.191</td>
<td>0.175</td>
<td></td>
</tr>
<tr>
<td>Scaled</td>
<td>0.989</td>
<td>0.975</td>
<td>0.953</td>
<td>0.966</td>
<td>1.001</td>
<td>0.980</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.211</td>
<td>0.193</td>
<td>0.205</td>
<td>0.184</td>
<td>0.197</td>
<td>0.177</td>
</tr>
<tr>
<td>Method</td>
<td>PC 2, $\sigma = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VD</td>
<td>0.226</td>
<td>0.231</td>
<td>0.199</td>
<td>0.212</td>
<td>0.214</td>
<td>0.228</td>
</tr>
<tr>
<td>Scaled</td>
<td>0.977</td>
<td>1.021</td>
<td>0.985</td>
<td>0.984</td>
<td>0.971</td>
<td>1.021</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.192</td>
<td>0.217</td>
<td>0.183</td>
<td>0.212</td>
<td>0.178</td>
<td>0.207</td>
</tr>
<tr>
<td>Method</td>
<td>PC 1, $\sigma = 0.1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VD</td>
<td>0.210</td>
<td>0.194</td>
<td>0.191</td>
<td>0.176</td>
<td>0.203</td>
<td>0.184</td>
</tr>
<tr>
<td>Scaled</td>
<td>0.988</td>
<td>0.975</td>
<td>0.953</td>
<td>0.966</td>
<td>1.001</td>
<td>0.980</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.210</td>
<td>0.193</td>
<td>0.205</td>
<td>0.185</td>
<td>0.196</td>
<td>0.178</td>
</tr>
<tr>
<td>Method</td>
<td>PC 2, $\sigma = 0.1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VD</td>
<td>0.225</td>
<td>0.232</td>
<td>0.199</td>
<td>0.213</td>
<td>0.213</td>
<td>0.229</td>
</tr>
<tr>
<td>Scaled</td>
<td>0.977</td>
<td>1.021</td>
<td>0.985</td>
<td>0.984</td>
<td>0.970</td>
<td>1.021</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.191</td>
<td>0.216</td>
<td>0.183</td>
<td>0.213</td>
<td>0.178</td>
<td>0.207</td>
</tr>
<tr>
<td>Method</td>
<td>PC 1, $\sigma = 1.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VD</td>
<td>0.235</td>
<td>0.230</td>
<td>0.225</td>
<td>0.211</td>
<td>0.225</td>
<td>0.220</td>
</tr>
<tr>
<td>Scaled</td>
<td>0.971</td>
<td>0.970</td>
<td>0.952</td>
<td>0.964</td>
<td>0.976</td>
<td>0.975</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.241</td>
<td>0.230</td>
<td>0.196</td>
<td>0.213</td>
<td>0.211</td>
<td>0.200</td>
</tr>
<tr>
<td>Method</td>
<td>PC 2, $\sigma = 1.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VD</td>
<td>0.211</td>
<td>0.264</td>
<td>0.204</td>
<td>0.256</td>
<td>0.207</td>
<td>0.266</td>
</tr>
<tr>
<td>Scaled</td>
<td>0.971</td>
<td>1.024</td>
<td>0.982</td>
<td>0.985</td>
<td>0.975</td>
<td>1.027</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.211</td>
<td>0.240</td>
<td>0.192</td>
<td>0.223</td>
<td>0.191</td>
<td>0.219</td>
</tr>
</tbody>
</table>

1 The distribution of domain lengths M is either Geometric(0.06) or Uniform(2, 80)
2 This is in reference to how eigenfunctions change as M increases. “Linear” refers to the eigenfunctions being linearly stretched over the domain length M. “NL” refers to the eigenfunctions being a “non-linearly” weighted combination of two distinct eigenfunctions where the weight is a function of M. Details are found in Section 4.3.1

### 4.4 Applications

We compare the performance of the three approaches, applying each method to two different applications. The first application uses the previously discussed
ICAP study, consisting of a daily recorded composite health score to track the progress of patients until death or discharge from the ICU. The second application presented is a case where the traditional functional methods applied to registered data could be defended scientifically; namely, data from hip-worn accelerometers that record movement and acceleration as patients stand up from a chair.

4.4.1 ICU Data

The ICAP study, the first application for which the three FPCA approaches are compared, consists of daily health status assessments for 512 patients in the ICU. Data has been made available in the refund package in R (Goldsmith et al., 2016). Health status is measured by SOFA scores until death or discharge of the patient, as described in Section 4.1. The duration of patient hospital stays in the study range from 2 to 170 days. The distribution of the domain length is quite skewed. Over 85% of observations have domain length $m \leq 30$ days. Figure 4.1-C shows the densities of domain length for subjects who died or were discharged, plotted on the log scale.

We apply variable-domain FPCA to the SOFA scores in the ICU. The heatmap in the upper left panel in Figure 4.5 displays the first principal component for $m \in (2, 170)$, ordered by domain length. In the upper right panel, we display the first principal component for domain lengths $m = 10, 60, 110,$ and $160$ days. The bottom panels of Figure 4.5 display the same results respectively for the second principal component.
Figure (4.5) Variable Domain PCs from ICU data: Left: Heatmaps of first and second principal component functions from vd-FPCA method across all domain lengths of the ICAP data from 2 to 170 days, ordered by domain length. Right: Selected examples of first and second principal components for $m = 10, 60, 110,$ and $160$ days. This shows the change in principal components as $M$ increases is more complicated than linearly scaling to the domain length. In parentheses above each plot is the percentage of total variance accounted for by the component.

The heatmaps in the two left panels indicate that the two first principal components are reasonably smooth; however there are two sudden shifts in the PCs around $m = 100$ and $140$. These shifts are likely related to the sparsity of the domain length values around these lengths of stay. Indeed, domain lengths greater than 100 days are fairly sparse and both of the sudden shifts for PC 2 in the lower left panel correspond to a gap where there are no subjects with domains of that
particular length (i.e. there are no subjects with hospital stays from approximately 80-100 days).

Figure 4.6 shows SOFA scores for subjects who have very high (black line) or very low (red line) scores for the first two PCs. Each plot shows the SOFA score trajectory and mean (dashed blue line) of two subjects at or around domain lengths $m = 10, 60, 110, \text{ and } 160 \text{ days}; \text{ results are shown for both PC 1 (top row) and 2 (bottom row). Subjects with the highest PC score were chosen among subjects with domain lengths within a certain neighborhood of } m. \text{ The width of the range increased with } m \text{ due to longer domain lengths not having as many subjects in their neighborhood (e.g. for } m = 10 \text{ days, only subjects with } m = 10 \text{ were examined, but for } m = 160 \text{ days, subjects having domain lengths from 150-170 were considered).}$

The four examples of the first two components on the right side of Figure 4.5 paint an intuitive picture of how the components change with $M. \text{ The components are not simply stretched over a longer domain or added to as } M \text{ increases. Instead, they display an increase in complexity that cannot be explained by scaling or trimming the principal components.}$

The plot of the first PC for $m = 10 \text{ in Figure 4.5 shows the variation between SOFA scores for subjects having domain length } m = 10 \text{ days in the ICU is mainly due to a shift in the overall level of their SOFA scores, with differences toward the end of the stay carrying slightly more weight. This is further demonstrated by the subjects with the highest and lowest PC 1 scores shown respectively with the black and red curves in Figure 4.6-A. Figures 4.5 and Figure 4.6-C show that for PC 1 at } m = 110, \text{ differences between subjects about 30-50 days into the stay are}
Figure (4.6) Example Participants by PC scores: Daily SOFA levels for subjects with hospital stays of length $m = 10, 60, 110,$ and $160$ days in each column. Black solid line shows SOFA for subject with a high (positive) PC score, red solid line shows subject with corresponding low (negative) score and dashed blue line shows the mean SOFA levels for each domain length $M$. The top row corresponds to PC 1, and the bottom row to PC 2. The plot for PC 2 for $m = 10$ days was excluded because the first PC accounts for more than 99% of the variation between subjects.

less important while the variation at the beginning and about two-thirds the way through the stay (around day 75) are more heavily weighted.

Thus, comparing the various PCs as $M$ increases, we see that the first principal components for the longer domain lengths of $m = 110$ and $160$ days are more complex than the relatively simple shifts up and down of the first PC for domain lengths $m = 10$ and $60$ days. Additionally, there could be health implications for the components that differ by domain length. For example, looking at the 21 subjects with domain length $m = 10$ days, 62.5% of those with high PC 1 scores died (the SOFA score increases, see subject shown in black in Figure 4.6-A)
which is significantly more than the 15.4% of those with low PC 1 scores that died ($p = 0.046$). Furthermore, if we just take the extreme scores, say everyone with scores lower than $-4$ or higher than 4, the probability of dying changes to 16.7% and 83.3% for the low and high scores respectively ($p = 0.0082$). However, if we examine the first PC for subjects with longer hospital stays, the same relationship does not hold. This demonstrates the importance of allowing the components to vary with the domain length in order to capture the increased complexity as the domain length increases.

The principal components estimated using the scaled and weighted methods described in Section 4.4 are shown in Figures 4.7-A and 4.7-B. If the first PC in Figure 4.7-A were multiplied by -1, it would be very similar to PC 1 in Figure 4.5 for $m = 10$. This suggests that the scaled data PCs are dominated by the characteristics of the shorter curves. To put this in perspective, 45% of subjects in the ICAP study spend less than 10 days in the ICU and 86% of subjects spend less than 30 days. Moreover, the results for the scaled domain do not seem to capture the main patterns of variation for higher lengths of stay. This is problematic, as individuals with longer lengths of stay exhibit more complex behaviors that could be of interest to survival.

The first five principal components estimated using the weighted method are more difficult to interpret and compare than the PCs from the first two methods. This is likely due to the fact that principal components are interpretable for a hypothetical subject with the maximum hospital stay, 173 days. However, in our application, 86% of the subjects spend less than 30 days in the ICU, which is before
the first minimum of PC 1 in Figure 4.7-B. Thus, before $t = 30$, the first principal component behaves similarly to the first components from the other methods. However, other characteristics are much harder to follow and characterize.

Our proposed method requires a slow increase in the number of components explaining 99% of the variability. Indeed, even up to 50-55 days (94% of the data) there are only 1 or 2 PCs that capture most of the observed variability. In Table 1, the number of PCs needed to account for at least 99% of the variation is shown as the length of the domain increases. At least 3 components are needed for subjects who stayed more than 55 days in the ICU. In contrast 6 PCs are necessary to capture the same amount of variability for the scaled version of the PCA. Moreover, the latter components are hard to interpret and are likely artifacts due to the stretching or compressing of the functional domain. Indeed, the higher order components both of the scaled and weighted PCA approaches resemble sinuses. This behavior
has little scientific meaning and is likely to be due to the lack of synchronization between important features of the data induced by time scaling.

Table 4.3: Number of principal components needed to account for at least 99% of the variation for corresponding domain lengths $M$.  

<table>
<thead>
<tr>
<th>Number of PCs</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain length $M$ (days)</td>
<td>1-22</td>
<td>23-55</td>
<td>56-78</td>
<td>79-157</td>
<td>158-173+</td>
</tr>
</tbody>
</table>

The variable domain method is more parsimonious than the alternative approaches and provides interpretable results with few principal components for every length of stay in the ICU. Indeed, for $M = 10$, one principal component accounts for more than 99% of the variation. For $M = 50$, the first component accounts for 81%, and the first 2 components account for more than 99% of the variation. In contrast, the first two principal components account for only 94% and 80% of the variation for the scaled and weighted methods respectively.

4.4.2 Accelerometry Data

We apply vd-FPCA to data collected from accelerometers worn on the hip by 20 subjects who repeatedly stood up from a chair during in-laboratory visits. Acceleration, measured in g’s ($9.81 \text{ m/s}^2$), was recorded by the accelerometer throughout the course of the standing movement. Multiple repetitions of the movement for each subject were separated and are treated as independent for the purposes of this analysis.

The time required for each of the 81 subject chair stands ranges from 0.55-2.30 seconds. Data were collected at 80 Hz, meaning each curve has between
50-180 observations. Moreover, roughly 60% of the data contain between 70 and 120 observations, demonstrating that the distribution of durations is more dense around longer domain lengths because there are multiple chair stands per person. More precisely, if a person stands up very slowly once, they are more likely to stand up slowly the next time.

This is a case when one could reasonably argue that the process of standing up from a chair for each person contains the same physiological phases. The movement is similar enough across subjects that traditional methods like scaling may have a more valid scientific foundation and perform better than in situations like the ICAP study, where domain lengths vary more and the functional observations are less scalable. Thus, one could argue that it could make sense to stretch or compress all accelerometry data to the same interval. We are not arguing in favor of doing this here, we just point out that this is a reasonable argument for the accelerometry data and an unreasonable one for the ICU data.

The interpretation of the principal components obtained via vd-FPCA from the accelerometry data, shown in Figure 4.8, is much the same as in the ICAP data example. The two heatmap plots show the first (top-left) and second (bottom-left) principal components over the observed domain, ordered by domain length. The plots on the right show specific examples of the PCs at $m = 0.7, 1.2, 1.7, \text{ and } 2.2$ seconds. The components are more smooth than in the ICAP study, likely due to the distribution of domain lengths not being as sparse in some areas as in the ICAP study.
Figure (4.8) Variable Domain PCs from Accelerometry Data: Left: Heatmaps of first and second principal component functions from vd-FPCA method across all domain lengths of the accelerometry data from 0.5 to 2.5, ordered by domain length. Right: Selected examples of first and second principal components for \( m = 0.7, 1.2, 1.7, \) and 2.2 seconds. In parentheses above each plot is the percentage of total variance accounted for by the component. This shows that change in principal components as \( M \) increases is more complicated than linearly scaling to the domain length.

For domains of length \( m = 0.5 \) to about 1.0 second in the upper-left heatmap plot of Figure 4.8, PC 1 seems to agree with the initial assumption that the principal component function is the same, just stretched or scaled over a larger domain. However, around \( m = 1.0 \) and 1.5 seconds, there is a change in the shape of the component, followed by another less dramatic shift around \( m = 2 \) seconds where the dip in the function shifts to earlier in the function. The shift is also evident in the second row of plots on the right for PC 1 where \( m = 1.7 \) and 2.2 seconds. The location of the minimum moves from \( t = 1.5 \) for \( m = 1.7 \)
to about 0.8 for $m = 2.2$ seconds. PC 2, as shown in the lower left section of Figure 4.8, conforms even less to the expected scaling over the varying domain of the component function than PC 1. Although the nature of the data would suggest scaling the data may be more appropriate, utilizing a variable domain approach can provide unique information that could be lost using the scaling approach.

Figure 4.9 shows the acceleration for subjects who have very high (black line) or very low (red line) scores for the first two PCs. The mean (dashed blue line) and acceleration trajectory of two subjects during the chair stand is shown for domain lengths $m = 0.7, 1.2, 1.7, \text{ and } 2.2$ seconds; results are shown for both PC 1 (top row) and 2 (bottom row). Subjects with the highest PC score were chosen among subjects with domain lengths within a neighborhood of $m$, typically around $\pm 0.1$ seconds. This plot is similar in nature to Figure 4.6.

From the plots relating to PC 1 in Figure 4.8 and Figure 4.9, we can see that the largest source of variation in subjects can mainly be attributed to one of two sources, depending on the domain length for the subject. The first is how fast or slow subjects accelerate to reach their peak acceleration, and the second is how quickly they stabilize after standing and decrease their acceleration. PC 1 for $M = 0.7$ and 1.7 seconds is related to subject stabilization after peak acceleration, whereas for $m = 1.2$ seconds, PC 1 is related to the speed of acceleration for the first portion of the movement. The heatmap for PC 1 in Figure 4.8 shows the principal component changes again for domain lengths of 2.2-2.5 seconds when the component is weighted positively for the entire middle section of the chair stand.
Subjects With High/Low Scores for PC’s 1 and 2

Figure (4.9) Example Participants by PC scores: Acceleration for subjects standing from a chair. Subjects with high (black solid line) and low (red solid line) corresponding PC scores are compared at domain lengths $m = 0.7, 1.2, 1.7,$ and $2.2$ seconds. The mean for each domain length is shown in blue (dashed line). Plots on the top row (A-D) correspond to PC 1 and plots on the bottom row (E-H) correspond to PC 2.

These results could indicate that depending on the duration of the chair stand, subjects can be discriminated by how quickly they reach peak acceleration or how quickly they can stabilize. Since the length of time it takes for a subject to stand and sit on a chair multiple times has been shown to be related to the risk of fall in older individuals (Buatois et al., 2008; Buatois et al., 2010), further investigation may be helpful to determine the strength of any correlation between PC scores based on the domain lengths and the risk of fall.

The principal components obtained from applying the scaled and weighted methods are shown in Figure 4.10. The first PC from the scaled method indicates the largest proportion of variation of the data is from the middle third of the
movement, likely due to differences in the height of the peak acceleration. This is similar to PC 1 for $m = 2.2$ seconds which may indicate that subjects having domain length around $m = 2.2$ seconds contribute more to the variation, even though only about 11% of the data has domain length greater than 2 seconds. PC 2 shows that the next largest proportion of variation from the stretched data is from having greater than average acceleration at the beginning of the chair stand and lower than normal acceleration in the latter portion of the movement. This is likely due to differences in when subjects reach their peak acceleration, similar to PC 2 for $m = 1.2$ and 2.2 seconds.

The first PCs from the weighted method are similar to that from the scaled method but have a different interpretation. The PCs of the scaled method are interpreted over the relative time of the movement compared to the absolute time for the weighted method. For example, the first PC from both methods has high
positive weights in the middle of the curve for the scaled method; this means that most of the variation between subjects occurs 1/3 to 2/3 the way through the motion of standing from a chair. For the weighted method, it means that most of the variation occurs from about 1 to 1.5 seconds into the motion, whether the subject takes 0.75 seconds to stand (meaning there is no recorded data from 1 to 1.5 seconds for that subject) or 2.0 seconds.

For \( m = 0.7 \), the first PC accounts for 63% of the variation, the first two PCs account for 89% and the first 3 PCs account for more than 99% of the variation. Principal components for \( m = 1.2 \) and 1.7 seconds account for about the same amount of variation as \( m = 0.7 \), but for \( m = 2.2 \) seconds the percentages are a little lower (58%, 81%, and 91% for 1, 2 and 3 PCs respectively). However, they are still higher than the standard methods used for comparison. For the scaled and weighted methods, the first PC accounts for 47% and 56% respectively, the first two PCs together account for 75% and 76%, and the first three PCs account for 88% and 87%, respectively.

<table>
<thead>
<tr>
<th>Table 4.4: Number of principal components needed to account for at least 99% of the variation for corresponding domain length ( M ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of PCs</td>
</tr>
<tr>
<td>Domain length ( M ) (sec.)</td>
</tr>
</tbody>
</table>

Figure 4.10 shows the PCs needed to account for at least 95% of the variation for each method; accordingly, the number of PCs is 8 for the scaled and 5 for the weighted methods, respectively. By contrast, the number of components needed using the variable domain method is shown for different domain lengths in Table 4. For shorter domain lengths, less than 1.43 seconds, only two or three PCs are
needed. Four or five are needed for the longer domain lengths, fewer than the scaled or weighted methods for all but the subjects having domain lengths greater than 2 seconds.

4.5 Discussion

The variable domain functional principal component analysis compares favorably to both the scaled and the weighted approaches. The scaled method implies that each curve is essentially similar across all percentiles of the curve, i.e. what is happening for the first 25% of the domain is the same regardless of whether it takes 3 days or 100 days to collect the data. This is arguably a reasonable assumption for the accelerometry study, where all subjects stand from a chair. Each subject performs a similar motion and data is observed until completion, regardless of the time to completion. However, for the ICU data, this assumption is invalid and limits study of the principal sources of variation across different domain lengths.

The weighted method implies that for any time $t = T$, the behavior of the curve at $T$ is the same for each curve. For example, at $t = 2$, the behavior of the function is the same regardless of whether the person stays in the hospital for 3 days or for 150. We have found the weighted method to be the least intuitive and hardest to explain and defend in a scientific context for these examples. Indeed it makes little sense to predict the SOFA score 100 days after a subject has either died or left the hospital, or to predict the acceleration 1 second after the subject is already at rest after standing up from a chair. However, there may be cases where this approach
may be reasonable, such as when subjects are observed over a fixed follow-up period and are either lost to follow-up, or have missing visits.

One of the major benefits of the variable domain approach is the increased flexibility of the principal component functions. The shape and number of components can vary according to the domain length and data complexity. In both study examples, the first few principal components were quite different for short domains compared to those corresponding to longer domains. For situations where the nature or shape of the principal component functions is important as the domain varies, this approach can provide practical and interpretable results that can be extended with additional models. For example, many of the principal components in the accelerometry study are related to subjects who increase and decrease their acceleration more quickly than others depending on how long they take to stand up. The interaction of PC scores and domain length could be used to explore relationships with various health outcomes such as whether older individuals who take longer to slow down their acceleration are more or less at risk for a fall or death.

In addition, this novel approach typically uses fewer components to explain a greater proportion of variation in the data than either of the alternative methods. This is particularly evident for observations with relatively shorter domain lengths. The parsimonious nature of the variable domain approach holds even in a non-traditional setting like standing from a chair where scaling the data to a singular domain makes intuitive sense. In both cases, we observed that the longer the domain, the more complex the underlying patterns of variation became. In turn,
this required a larger number of components to account for a given percentage of the variation in the data as the length of the domain increased. We contend that the variable approach, which conducts dimensionality reduction on the natural scale of the data, provides tailored insights into the variability of the data as a function of the domain. Once the initial hurdle of working with domain-dependent PCs is passed, the results are quite intuitive and easy to understand.

As we showed in the simulation results, the distribution of the domain lengths, noise levels, and the manner in which the eigenfunctions vary across domain lengths each have a large impact on performance for each approach. The variable domain approach tends to perform best in situations where the distribution of domain length is skewed or the eigenfunctions are more complicated than a linear rescaling across the domain length. Thus, it is important to take this into account when selecting an approach.

One limitation of this approach is the computation time due to fitting a trivariate smoother on so many observations. It can take over an hour to fit the covariance function on the accelerometry data which has a relatively small number of subject curves (81) even though the observations are fairly dense, having about 50 - 200 observations per curve. Thus, scaling up the current approach to truly large datasets is currently not feasible; however, alternative approaches without this limitation are being studied.

In this analysis, we focus on the case where all subjects have a fixed starting point and have variable ending points. One possible extension of this method in the future would be to have variable starting and ending points with a landmark
time in common for all subjects somewhere in the middle, say at \( t = 0 \). We are currently working on approaches to handle this type of scenario and extend this methodology to other situations as well.

Despite the limitations due to computation time, vd-FPCA is a viable alternative option for determining key patterns of variation in variable domain data, particularly in applications where health outcomes are potentially related to the domain length of the functional data.
4.6 Appendix

Theoretically, there should be no difference between estimating non-parametrically a function \( f(x) \) or \( f\{T(x)\} \), where \( T(x) \) is a known transformation of the domain of the function \( x \rightarrow T(x) \). Since \( f(\cdot) \) is not specified, \( h = f \circ T \) is not specified and should be as easy to estimate non-parametrically as \( f(\cdot) \). Unfortunately, in theory and in practice, this is not the case and the transformation of the \( x \)-space leads to substantially different estimators. To be concrete, consider the case when one is interested in estimating the smooth function \( f(\cdot) \) from pairs of observed data \((x_i, Y_i)\). A standard problem to solve in this case is to minimize the penalized criterion

\[
\sum_{i=1}^{n} \{Y_i - f(x_i)\}^2 + \lambda \int \{f''(x)\}^2 dx ,
\]

where \( \lambda \) is a smoothing parameter applied to the integral of the square of the second derivative of \( f(\cdot) \). Of course other penalties could be used, but here we focus on the standard one. The following ideas can easily be generalized to other penalties. Consider the case when the \( x_i \)'s are transformed into \( z_i = T(x_i) \). In this case, one estimates a function \( g(\cdot) \) such that the penalty becomes

\[
\sum_{i=1}^{n} \{Y_i - g(z_i)\}^2 + \lambda \int \{g''(z)\}^2 dz ,
\]

or, re-writing in terms of \( x \)

\[
\sum_{i=1}^{n} \{Y_i - g(T(x_i))\}^2 + \lambda \int \{g''\{T(x)\}\}^2 dx .
\]
If we wanted the exact same smoother then we would need to penalize the square of the second derivative of $g \circ T$, which is equal to

$$g''\{T(x)\}\{T'(x)\}^2 + g'\{T(x)\}T''(x) \neq g''\{T(x)\}.$$  

The equality holds if the transformation $z_i = T(x_i)$ is linear, but does not typically hold for other types of transformations. In short, transforming the function domain, then smoothing on the new domain and transforming back to the original domain is not equivalent to smoothing on the original domain unless the transformation is linear. One could obtain the exact same estimate if the penalty term were changed to match the transformation, but almost never is this the case in practice. This is not only true in theory, but in practice as well. Below we show how the curve estimate can change dramatically via the domain transformation and the back transformation after smoothing.

To illustrate this point, consider the class of functions

$$m(x, j) = \sqrt{x(1-x)} \sin \left( \frac{2\pi(1 + 2^{(9-4j)/5})}{x + 2^{(9-4j)/5}} \right)$$  

for $x \in [0, 1]$ (4.8)

and the class of transformations of the domain $T(x) = x^\alpha$ for $\alpha > 0$. Figure 4.11 shows how the function $m(x, 6)$ changes as the domain is transformed by $T(x) = x^\alpha$ for four different levels, $\alpha = (0.1, 0.5, 2, 10)$. When $\alpha > 1$, the points are shifted to the left and compacted to be close to 0, but when $\alpha < 1$, the points are shifted to the right and compacted close to 1. This class of functions has been used extensively to illustrate adaptive smoothing and is considered to be a difficult class of functions to fit.
Figure (4.11) Non-linear Transformations: The black dashed line shows the original function \( m(x, j) \) from Equation (8) where \( j = 6 \). The red line and corresponding points show the same function with the transformed domain \( T(x) = x^\alpha \) for \( \alpha = (0.1, 0.5, 2, 10) \).

Figure 4.12 shows the result of smoothing the function on the original domain (shown in red) compared to the smoothed results of the transformation \( T(x) = x^\alpha \) for \( \alpha = 0.01 \) and 20 (shown in green and blue respectively). The original data is shown in gray. These results were obtained by fitting a cubic smoothing spline using the function smooth.spline in R (R Core Team, 2018). Each spline had 150 knots and the smoothing parameter was determined by generalized cross-validation. After fitting the data, everything is transformed back to the original scale. Note the substantial differences among these fits. Indeed, when the transformation is \( T(x) = x^{20} \) (shown in blue) the entire first part of the function is being missed, while the second part of the function indicates a strongly unsmooth pattern. When
the data is compacted to the left before smoothing, the first part of the function fits relatively well, but the second part still displays a strong under-smoothed pattern (red). After the transformation corresponding to $\alpha = 0.01$, the estimator performs exceptionally well, even though we are applying a standard smoother.

There are two important conclusions here: 1) transforming the x-space, smoothing using standard techniques, and re-transforming back to the original scale of the data is not invariant to non-linear transformations; 2) the more one understands the underlying structure of the x-space, the better the smoothing results; 3) transformation of the x-space should be coupled with the corresponding transformation of the penalty term on the unspecified smooth function. We would also like to highlight a completely unexpected finding. Indeed, these functions have been introduced in the literature as examples of hard functions to smooth and have been used extensively to highlight the power of adaptive smoothing. More precisely, methods have been proposed to either increase or decrease the number of knots for splines or allow for a different smoothing parameter as a function of the domain (Crainiceanu et al., 2007).

Thus, we are showing that for this class of functions, these approaches are actually not necessary and a careful transformation of the x-domain can provide extraordinary improvements in the smoothing approach without requiring advanced adaptive smoothing techniques. In other words, we have achieved adaptive smoothing using standard smoothing plus a simple transformation of the x-space. This further highlights that transforming the x-space can have substantial, unexpected, and strong effects on the smoothing estimators.
Figure (4.12) Smoothing with Non-linear Transformations: Smooth estimates of a function with and without transformation of the domain. Raw data is shown in dark gray and the smoothed data on the original domain is shown in red. The green and blue curves show the smooth estimates after transformation of the domain, $x^\alpha$ for $\alpha = 0.01$ and 20 respectively.
References


Chapter 5

Discussion and Conclusion

5.1 Predicting multiple sclerosis lesion development using longitudinal magnetic resonance imaging

In chapter 2 we have shown that in spite of the noise and other complexities in lesion voxel trajectories, we are able to produce high quality predictions up to a year in the future using MRI brain scans from just a few visits spaced out monthly. The dynamic linear models that were used were fast and easy to implement and interpret. Multiple models were fit and showed that models including historical information concerning voxel behavior was more accurate than models with only demographic, treatment, or nearby voxel information.

The image tools that we have presented have the potential to be a useful tool for clinicians and researchers. They help visualize and facilitate understanding of the trajectory of a lesion or voxel region which could help advise personalized treatment decisions. They provide quick access to in-depth analysis results for MS lesions in an interactive manner from a variety of perspectives.
We have also introduced a straightforward method of comparison of lesion behavior at the voxel level to pre-incidence levels. These developments demonstrate significant progress toward understanding lesion behavior and being able to predict if a lesion region will recover and how long it will take to do so.

5.2 Fragmentation in Ecological Momentary Assessment

In chapter 3, we have shown that fragmentation methods, measures of stability that historically have not been used with EMA data, can provide additional information beyond that of more traditional stability and variability measures, particularly concerning the characterization of stability for specific psychological states.

We have also showed that individuals with BPI disorder have less stability when inside their standard range of attention than controls. On the other hand, they exhibit greater stability and have more assessments that fall outside the standard range of attention. Individuals with MDD have less stability than controls when inside their standard range for attention and mood. The lack of significant differences in mood for those with BPI replicates clinical studies that highlight that attention is a more important psychological process in BPD than mood (Scott et al., 2016; Johnson, Gershon, and Starov, 2015).

The validity of the results obtained in this sample should be tested over longer periods of assessment, as well as with additional types of mental disorders. They are also dependent on the definition of the "standard state" and subject to influence from the sampling frequency and timing of assessments. Additionally, we did not
address directionality, namely, these fragmentation measures are not set up to determine whether differences in stability between mood disorder subgroups are related to high or low scores (i.e., if a subgroup is more stable when their mood and attention levels are either high or low). Future work will involve the extension of fragmentation measures to account for this across subgroups.

5.3 Variable Domains

In chapter 4 we have shown that variable domain functional PCA compares quite favorably to other approaches in a variety of ways. The variable domain approach performed best in scenarios where the domain length distribution was heavily skewed or if the eigenfunctions vary non-linearly with the domain length. We also have shown that the performance of the principal component analysis can widely vary due to noise levels in addition to distribution of domain lengths and their interaction with eigenfunctions; thus, it is important to take this into account when selecting an approach.

We have found that the variable domain approach typically uses fewer components to explain a greater proportion of variance, when compared with the alternative methods. This was particularly evident with the individuals having shorter domain lengths, even in settings where the distribution of domain lengths was narrow. The flexibility this approach demonstrated for eigenfunctions with different domain lengths is particularly attractive, facilitating tailored insights into the variability of the data as a function of the domain.
5.4 Conclusion

We have shown that it is important for the individuals performing statistical analyses to understand and be able to adapt to the unique circumstances of the current data and study design. Consequently, data exploration and visualization is an important step in any data analysis. We have demonstrated various ways of how this can been done in a variety of situations: 1) dynamically predicting voxel trajectories after lesion incidence 2) associating stability and fluctuations in mood and attention in EMA data with mood disorder subgroups, and 3) finding typical modes of variation for data with varying functional domains. Some of these examples can easily be extended to other types of data, others are more tailored to a very specific scenario, but all provide unique insights that facilitate reproducible data analysis and address concerns in each of their respective fields.
References


Jordan T. Johns
Ph.D. Candidate, Biostatistics • Johns Hopkins University Bloomberg School of Public Health
615 N. Wolfe St. • Baltimore, MD 21205
Phone: 801-830-2098 (c) • E-Mail: jordan.johns@jhu.edu

Research interests

• Wearable devices, longitudinal studies, electronic diaries, functional data analysis, structural brain imaging, personalized medicine

Education

• Ph.D., Biostatistics, Johns Hopkins University (2019), Baltimore, MD
• M.S., Statistics, Brigham Young University (2014), Provo, UT
• B.S., Statistics, Brigham Young University (2012), Provo, UT
  o Magna Cum Laude
  o GPA: 3.93/4.0
  o Minors in Mathematics and Spanish

Professional Experience

Pre-Doctoral Fellow, NIH, National Institute of Mental Health (2015-2018)

• Development of statistical methods for large community based study of rhythms of emotion, activity, and sleep as trait markers for individuals with the full range of mood disorders.
• Genetic Epidemiology Research Branch under Kathleen Merikangas.

Teaching Assistant, Johns Hopkins University (2015-2018)

• Master of Public Health Capstone Project
• Statistical Reasoning in Public Health I-II
• Statistical Methods in Public Health I-IV
• Methods in Biostatistics I-II


• Individual research project concerning the use of empirical Bayes estimators in pharmaceutical settings.
• Verification of statistical analyses, calculating sample sizes, determination of randomization schedules, and managing datasets according to CDISC standards.
Research Intern, Johns Hopkins University (2013)
- Worked with technology research group using graphical and functional data methods on accelerometer data to quantify and classify movements using unsupervised clustering methods. This research was further developed into Master’s Degree capstone project at BYU.
- Assisted in forming short course on structural MRI data analysis.

Statistics Department, Brigham Young University
- Teaching Assistant (2012–2014)
  o Survival Analysis
  o Statistical Methods for Research I-II
  o Probability Theory and Math Stat
- Graduate Assistant (2012)
  o Provided statistical support to faculty and graduate students in the BYU Consulting Center.
- Research Assistant (2011–2012)
  o Implemented survival analysis techniques in high dimensional and traditional settings such as Cox proportional hazards models and hierarchical clustering.

Special Education Research Assistant, Brigham Young University (2011–2012)
- Worked with faculty members to design methods in order to analyze teacher candidate performance.
- Managed data and prepared reports to demonstrate that the Special Education program meets CEC Standards in preparing teacher candidates for licensing.

Publications / Presentations


**Skills**

Language: English, Spanish

Document Preparation: Microsoft Word, Latex, RMarkdown

Software/Language: R, SAS, Matlab, Stata, SQL, C++, Shiny

Operating Systems: Windows, Mac

Updated May 2019