Functional Data Analysis and Machine Learning for High-Dimensional Structured Data

Ángel García de la Garza

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Functional Data Analysis and Machine Learning for High-Dimensional Structured Data

Ángel García de la Garza

This thesis pertains to the uses of Functional Data Analysis and Machine Learning when analyzing high-dimensional structured datasets. The theme that motivates the first two chapters is the development of dimension-reduction methods in the context of functional data to advance the understanding of in-vivo measurements of neural-spike data. The last chapter addresses the analysis of survey data using machine learning techniques to identify novel risk factors for suicide in the general population.

The first chapter of this thesis, "Adaptive Functional Principal Component Analysis," provides a novel method for adequately capturing modes of variation in data exhibiting sharp changes in smoothness. Our work integrates a novel scatterplot technique that adaptively smooths latent functions estimated in an FPCA framework. We are motivated to identify coordinated patterns of brain activity across multiple simultaneously-recorded neurons during motor behavior to understand the dynamics between the brain and dexterous movement. Our proposed method adequately captures the underlying biological mechanisms in our experiment, offering interpretable activation patterns when compared to standard approaches.

The second chapter of our dissertation develops statistical procedures to compare the eigendecomposition from two samples of functional data. We first introduce appropriate tests for both independent and paired functions. We are motivated to test whether activation patterns in the motor cortex hold constant when a mouse performs a reaching movement repeatedly. We test all

pairwise comparisons across trials and compare the distribution of the p-values against the distribution under the null. Our results suggest trial-to-trial variation in the latent activation patterns that can't be attributed to sampling noise. Our results can inform future methodology for deriving activation patterns from noisy neural spikes.

The last chapter of this dissertation dives into applying Machine Learning Techniques to analyze survey data. We use the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) survey to identify novel risk factors for suicide attempts in the general population. Our analysis uses a Balanced Random Forest (BRF) approach and incorporates extreme class imbalance and survey architecture into the algorithm. We extend prior research focusing on high-risk clinical samples by identifying risk factors for suicide attempts in the general population. Our work identifies risk variables that can help guide clinical assessment and the development of suicide risk scales.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I want to thank my advisor, Jeff Goldsmith. I am grateful for your support and guidance throughout the last five years. You've taught me all the fundamentals of FDA and Data Science, and your advice helped me grow as an academic.

Christine Mauro, I am grateful for your mentorship, guidance, and friendship. You've helped me navigate graduate school. Todd Ogden, thank you for your mentorship and feedback throughout the years at FDAWG. Melanie Wall, I am thankful for your mentorship and for giving me the chance to work in the field of Mental Health Statistics. Britton Sauerbrei, thank you for giving me the chance to work in Neuroscience and for patiently helping me improve my understanding of the field. I am also thankful to Carlos Blanco and Mark Olfson for their guidance.

I am grateful to Yifei Sun, for your thoughtful comments and suggestions throughout the years and for teaching me the fundamental of Machine Learning. Seonjoo Lee, thank you for your insightful comments and suggestions as a member of my committee.

Thank you to the Department of Biostatistics at Columbia for supporting me financially. I am also grateful to Katy Hardy, Georgia Andre, and Jessica Jimenez for helping me navigate hiring and payroll. I am grateful to Justin Herrera for your support throughout the years, for allowing me to be part of BEST, and for all your help running the Computing Club.

I am grateful to Taki Shinohara, Ted Satterthwaite, Kosha Ruparel, Raquel Gur, Ruben Gur, Andrea Troxel, and Linda Zhao for your mentorship.

I thank Pia Figuerola, Melissa Vela, Damani Clarke, Vedika Luharuwalla, Elaine Song,

Rebecca Venetianer, Marco Herndon, Mabel Oviedo, Daniel Saenz, and Gus Leinbach for their invaluable friendship through the years. I also thank my peers, Soohyun Kim, Muhire Kwizera, Julia Wrobel, Maddie Stoms, Amy Pitts, Abby Zhao, Margaret Gacheru, and Melanie Mayer, for their friendship and help in research and classes.

My most important thanks go to my parents for their love and support throughout my studies at Columbia and the last 11 years in the US. I also thank my sister, Alejandra, for all her love and advice throughout the years.

Chapter 1: Introduction

The overall theme of this thesis is the analysis of high-dimensional structured datasets. The first section of the thesis explores this idea in the context of dimension reduction for Functional Data Analysis. The first chapter presents a novel statistical methodology to derive latent functions in data exhibiting sharp changes in smoothness by using adaptive scatterplot smoothing in the estimation of Functional Principal Component Analysis. The second presents a novel test to compare latent functions in two samples of functional data. This work is motivated by understanding the intricacies of the biological basis of behavior by analyzing brain activation patterns derived from in-vivo measurements of neural-spike data. The second section addresses the analysis of survey data using machine learning in the context of mental health data science. Our algorithm integrates common features in psychiatric surveys, such as extreme class imbalance, informative missingness, and complex survey architecture. This work is motivated by the need to identify novel risk factors for suicide in non-clinical samples, as the general population is understudied and represents a large proportion of suicide attempts.

The first chapter of this thesis, "Adaptive Functional Principal Component Analysis," provides a novel method to adequately capture modes of variation in functional data when the underlying smoothness varies over the domain. We propose a new adaptive scatterplot smoothing technique that is scalable to high-dimensional data and integrate it to adaptively smooth latent functions estimated in an FPCA framework. We are motivated to identify coordinated patterns of brain activity across multiple simultaneously-recorded neurons during motor behavior to understand the dynamics between the brain and dexterous movement. Our proposed method adequately captures the underlying biological mechanisms in our experiment, offering interpretable activation patterns versus standard approaches. Our simulations mimic functional data with sharp changes in smoothness. Results show that our method is better suited for modeling these datasets than standard approaches.

We develop accompanying publicly available software for our proposed methodology.

The second chapter of our dissertation develops statistical procedures to compare the eigendecomposition from two samples of functional data. We first introduce an appropriate test when the observations of both groups are entirely independent and extend it to the case of paired functions. Our procedure tests the covariance matrix of the FPCA scores rather than comparing the eigendecomposition directly. Simulation studies suggest our proposed methods have significantly higher power when compared to other leading tests across various scenarios. We are motivated to test whether activation patterns in the motor cortex hold constant when a mouse performs a reaching movement repeatedly. We test all pairwise comparisons across trials and compare the distribution of the p-values against the distribution under the null. Our results suggest trial-to-trial variation in the latent activation patterns that can't be attributed to sampling noise. Our results provide insightful new knowledge about how to derive activation patterns from noisy neural spikes.

The last chapter of this dissertation dives into applying Machine Learning Techniques to analyze survey data. We use the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) survey to identify novel risk factors for suicide attempts in the general population. Our analysis uses a Balanced Random Forest (BRF) approach and incorporates extreme class imbalance and survey architecture into the algorithm. We also leverage the extensive assessment instrument that captures information not routinely available in Electronic Health Records (EHR) or administrative data. The majority of previous work in the literature focuses on high-risk clinical samples. However, a third of people attempting suicide do not receive mental health treatment. Our work extends suicide risk research by identifying risk factors in the general population. We extend prior research focusing on identifying risk factors for suicide attempts in the general population. Our work identifies risk variables that can help guide clinical assessment and the development of suicide risk scales.The overall theme of this thesis is the analysis of high-dimensional structured datasets. The first section of the thesis explores this idea in the context of dimension reduction for Functional Data Analysis. The first chapter presents a novel statistical methodology to derive latent functions in data exhibiting sharp changes in smoothness by using adaptive scatterplot

smoothing in the estimation of Functional Principal Component Analysis. The second presents a novel test to compare latent functions in two samples of functional data. This work is motivated by understanding the intricacies of the biological basis of behavior by analyzing brain activation patterns derived from in-vivo measurements of neural-spike data. The second section addresses the analysis of survey data using machine learning in the context of mental health data science. Our algorithm integrates common features in psychiatric surveys, such as extreme class imbalance, informative missingness, and complex survey architecture. This work is motivated by the need to identify novel risk factors for suicide in non-clinical samples, as the general population is understudied and represents a large proportion of suicide attempts.

The first chapter of this thesis, "Adaptive Functional Principal Component Analysis," provides a novel method to adequately capture modes of variation in functional data when the underlying smoothness varies over the domain. We propose a new adaptive scatterplot smoothing technique that is scalable to high-dimensional data and integrate it to adaptively smooth latent functions estimated in an FPCA framework. We are motivated to identify coordinated patterns of brain activity across multiple simultaneously-recorded neurons during motor behavior to understand the dynamics between the brain and dexterous movement. Our proposed method adequately captures the underlying biological mechanisms in our experiment, offering interpretable activation patterns versus standard approaches. Our simulations mimic functional data with sharp changes in smoothness. Results show that our method is better suited for modeling these datasets than standard approaches. We develop accompanying publicly available software for our proposed methodology.

The second chapter of our dissertation develops statistical procedures to compare the eigendecomposition from two samples of functional data. We first introduce an appropriate test when the observations of both groups are entirely independent and extend it to the case of paired functions. Our procedure tests the covariance matrix of the FPCA scores rather than comparing the eigendecomposition directly. Simulation studies suggest our proposed methods have significantly higher power when compared to other leading tests across various scenarios. We are motivated to test whether activation patterns in the motor cortex hold constant when a mouse performs a reaching

movement repeatedly. We test all pairwise comparisons across trials and compare the distribution of the p-values against the distribution under the null. Our results suggest trial-to-trial variation in the latent activation patterns that can't be attributed to sampling noise. Our results provide insightful new knowledge about how to derive activation patterns from noisy neural spikes.

The last chapter of this dissertation dives into applying Machine Learning Techniques to analyze survey data. We use the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) survey to identify novel risk factors for suicide attempts in the general population. Our analysis uses a Balanced Random Forest (BRF) approach and incorporates extreme class imbalance and survey architecture into the algorithm. We also leverage the extensive assessment instrument that captures information not routinely available in Electronic Health Records (EHR) or administrative data. The majority of previous work in the literature focuses on high-risk clinical samples. However, a third of people attempting suicide do not receive mental health treatment. Our work extends suicide risk research by identifying risk factors in the general population. We extend prior research focusing on identifying risk factors for suicide attempts in the general population. Our work identifies risk variables that can help guide clinical assessment and the development of suicide risk scales.

Chapter 2: Adaptive Functional Principal Component Analysis

2.1 Introduction

Functional data analysis (FDA) is concerned with settings where observations made on study units are functions measured over time, space, or another continuum. Methods for analyzing such data borrow information across adjacent points in the functions' domain and therefore differ from multivariate approaches [1]. For example, functional principal component analysis (FPCA) is a dimension reduction technique that identifies a set of orthogonal functional principal components (FPCs) that are continuous and smooth. A central consideration in FDA is how to model smoothness most appropriately when conducting an analysis. Whether through a basis expansion, the structure of smoothness-enforcing penalties, or some other mechanism, FDA methods, including FPCA, implicitly assume similar degrees of smoothness across the functional domain. When the underlying smoothness in the data fluctuates, this will lead to models that under and over-smooth over different sections of the data domain.

Samples of curves that exhibit locally-varying degrees of smoothness arise regularly. In each trial in the experiment that motivates our work, a trained mouse reaches for a food pellet after hearing an auditory cue while continuous measurements of spike activity in 25 neurons on the motor cortex are recorded using silicon probes [2]. Before the cue, the mouse's motor cortex is at rest; the auditory cue triggers an immediate response in the motor cortex and, subsequently, a voluntary reach. In the later stages of the reach, neural activation declines slowly and smoothly. Figure 1 summarizes our data. Panel A1 shows binary neural activation across the reaching movement recorded in 10ms windows for each of 157 trials in four representative neurons. We average across trials to obtain each neurons' typical activation during the reaching experiment. Panel A2 shows the trial-averaged neuron-specific means measured in spikes per second for the same four neurons.

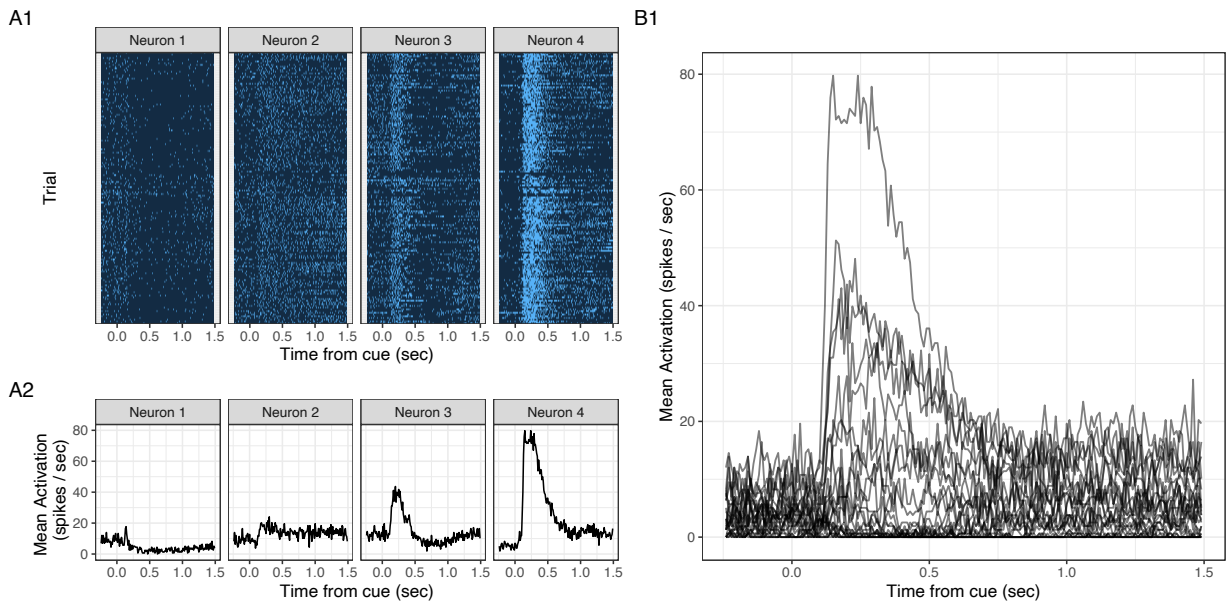


Figure 2.1: Panel A1 displays lasagna plots of the activation of four example neurons over a 1.75 seconds interval beginning 0.25 seconds before an auditory cue and 157 trials. Light blue indicates that the neuron is active. Panel A2 displays the average activation of the same four neurons. Panel B plots the average activation of the 25 neurons in our sample.

The resulting neural processes exhibit sharp changes in activity in response to the auditory cue and periods of smoothly changing activation during the reach. Panel B1 shows the trial-averaged activation for all 25 neurons, which will be the focus of our analysis.

Our scientific goal is to identify the activation patterns that emerge across neurons during voluntary motor behavior in order to better understand the involvement of the motor cortex in skilled movements. Patterns derived using state-of-the-art methods for dimension reduction fail to capture the non-constant smoothness in these data and thus do not reflect the underlying biological behavior. We propose an innovative approach to dimension reduction for functional data in which the level of smoothness varies locally. We first develop a new technique for locally adaptive scatterplot smoothing based on the Adaptive ridge penalty and then incorporate that into the estimation of functional principal functional components (FPC) using a penalized likelihood framework. Simulations indicate that our proposed Adaptive FPCA method outperforms competing approaches when the data generating mechanism includes non-constant degrees of smoothness. Although it

is not necessary for our motivating data, our implementation allows sparse and irregular grids for observed functional data. Applying this method to our motivating data yield to interpretable activation patterns across the motor cortex, clearer scientific conclusions, and robust fits to observed curves.

The rest of this manuscript is organized as follows. Section 2 provides a review of the relevant literature. Section 3 contains subsections on the penalized likelihood approach to FPCA, our method for adaptive scatterplot smoothing and the adaptive FPCA model specification. Section 4 presents simulation, and Section 5 demonstrates the application of our method's to our motivating neuron spike data. We close with a discussion in Section 6.

2.2 Literature Review

Our contributions build on prior work in functional principal components analysis, adaptive scatterplot smoothing, and adaptive ridge penalties; we review the relevant literatures in Sections 2.1, 2.2, and 2.3, respectively.

2.2.1 Functional Principal Component Analysis

Because we are primarily interested in the role of smoothing in dimension reduction, we focus our review on existing approaches to smoothing in FPCA. FPCs are frequently obtained through an eigendecomposition of the covariance operator of functional observations $Y(t)$, defined as $\Sigma(u, v) = \text{Cov}\left(Y(u), Y(v)\right)$ [3]. Early approaches to FPCA smoothed observed curves before estimating the covariance operator [4] or implemented a non-functional PCA and smoothed the resulting components to obtain FPCs [5, 6, 7]. More recently, it has been common to smooth an empirical covariance estimated from observed data and then decompose the result. Examples of this general approach include bivariate kernel or kernel-based approaches [8, 9, 10]; penalized tensor product splines [11, 12]; and fast bivariate P-splines [13, 14].

Methods based on probabilistic principal component analysis [15] estimate FPCs by maximizing a likelihood rather than estimating, smoothing, and decomposing a covariance operator. As

a result, they may be appealing for data observed over sparse or irregular grids or when the dimension of the observation grid makes smoothing an empirical covariance computationally challenging. Probabilistic FPCA methods include the latent-factor approach for Gaussian data by [16]; the variational Bayesian approach for binary and count data developed by [17]; and the Bayesian generalized multilevel FPCA extension developed by [18]. These approaches estimate FPCs directly and often include explicit penalties to enforce smoothness on the results. In contrast to our proposed methods, however, neither covariance-based nor probabilistic approaches allow for locally-varying degrees of smoothness.

2.2.2 Adaptive Scatterplot Smoothing

We next discuss techniques for adaptive scatterplot smoothing in non-functional settings. Scatterplot smoothing considers observations $\{(t_i, y_i) : i = \{1, \ldots, I\}\}$ and focuses on estimating $E[y] = f(t)$ as a smooth function of t. The goal is to estimate $f(\cdot)$ in a way that balances the goodness of fit to the data against the complexity of the resulting $\hat{f}(\cdot)$. A common approach is to penalize the outcome likelihood by using the integrated squared second derivative $\lambda \int f''(t)^2 dt$ with a smoothing parameter λ that tunes the contribution of the penalty term. The literature on scatterplot smoothing is too vast to thoroughly review here, and instead we will focus narrowly on spline-based methods. As a starting point, we assume that $f(\cdot)$ is expressed as a set of spline basis functions and respective coefficients. The ridge penalty [19, 20] is a useful tool to implement non-adaptive scatterplot smoothing, as the integrated squared second derivative penalty is can be expressed as an L_2 penalty on the spline coefficients [21, 22, 23]. This relationship underlies many techniques in scatterplot smoothing and functional data analysis. For the purposes of this manuscript we assume cubic B-splines and a second derivative penalty. Other spline basis are possible, and all our methods are implemented using B-splines or truncated polynomials.

A limitation of the non-adaptive smoothing penalty is that it penalizes the overall smoothness of the estimated function. When the degree of curvature in $f(t)$ varies over t, standard approaches locally under or over-smooth [24]. This can be addressed by the penalty function $\lambda(t)$

over the domain of the data rather than a tuning constant λ [25, 26, 27, 28]. The adaptive penalty $\int \lambda(t) f''(t)^2 dt$ assigns local weight to the smoothness of the fit $f''(t)^2$. [29] implemented adaptive scatterplot smoothing by finding a set of penalty constants using the GCV criterion. [30] proposed a hierarchical model, in which spline coefficients have priors with unique variances, fitted using Markov chain Monte Carlo (MCMC). [31] developed a fast implementation of adaptive smoothing using a similar hierarchical model and an iterative algorithm. [32] estimates $\lambda(x)$ as a step-function fitted using an AIC-like criterion. Other alternatives that do not use the penalty function $\lambda(x)$ include the fused lasso adaptive model procedure (FLAM), in which the estimated functions are piece-wise constant with adaptively chosen knots [33] or the approach undertaken using neighbor fused lasso by [34].

2.2.3 Adaptive Ridge Penalty

Section 2.3.2 proposes a novel technique that casts the adaptive second derivative penalty as an adaptive ridge penalty on spline coefficients. For coefficients $\{\beta_p : p \in \{1, \ldots, P\}\}$, to be estimated, a standard ridge penalty including the tuning parameter λ takes the form $\lambda \sum_{p=1}^{P} \beta_p^2$. This regularizes the coefficients and has a closed form solution for fixed values of λ and outcomes that have a Gaussian distribution [19, 20].

The Adaptive ridge (AR) is a modification of the ridge that assigns a different tuning parameter to each of the coefficients through the penalty $\sum_{p=1}^{P} \lambda_p^2 \beta_p^2$ [35, 36]. This penalty is "adaptive" in the same sense as the adaptive LASSO, in that each coefficient has a unique tuning parameter or weight; it is not "adaptive" in the same sense as adaptive scatterplot smoothing. AR penalties are implemented using iterative algorithms that alternate between updating the tuning parameters $\{\lambda_p : p \in \{1, \ldots, P\}\}$ based on the previous estimates of the coefficients based on the previous estimates of the coefficients and re-estimating $\{\beta_p : p \in \{1, \ldots, P\}\}$ given these updated weights [37, 38]; both steps have computationally efficient closed-form solutions. Our work on adaptive-scatterplot smoothing and adaptive FPCA casts the adaptive squared second derivative penalty in terms of a computationally inexpensive AR penalty, and estimates spline coefficients and tuning

parameters as alternate steps in an iterative algorithm.

2.3 Methods

We propose new methods to identify motor cortex activation patterns from neural spike data that exhibit sharp changes in response to a cue but are smoothly varying elsewhere in the observation. First, we briefly review a likelihood-based method for FPCA that enforces the same degree of smoothness over the functional domain and therefore is not suited for our data. We next introduce our novel approach to adaptive scatterplot smoothing, and then our FPCA methods that are able to capture sharp changes in smoothness in the patterns that underlie observed data.

2.3.1 Statistical Framework

Define $X_i(t)$ be a set of functions measured over $t \in \mathcal{T}$ for observation $1 \leq i \leq I$ with common mean $\mu(t)$ and covariance operator $\Sigma(u, v) = \text{Cov}[X(u), X(v)]$. Mercer's theorem provides a decomposition of the covariance operator based on eigenvalues and eigenfunctions; a Karhunen-Loeve (KL) expansion of functions $X_i(t)$ using these is given by,

$$E[X_i(t)] = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t) \tag{2.1}$$

where $\mathbf{\Phi}(t) = \{\phi_k(t) : k \in \mathbb{Z}^+\}$ are orthonormal eigenfunctions, $\boldsymbol{\eta} = \{\eta_k : k \in \mathbb{Z}^+\}$ are the corresponding eigenvalues, and scores $\xi_{ik} = \int_0^1 [X_i(t) - \mu(t)]\,\phi_k(t)dt$ are uncorrelated random variables with mean zero and variance η_k. In the analysis of a sample of curves, the expansion in (2.1) is truncated to retain the first K eigenfunctions. Although motivated by the decomposition of a covariance operator, all terms in the truncated KL expansion can be estimated directly using a likelihood-based approach.

In real data settings, we observe $Y_i(t) = X_i(t) + \epsilon_i(t)$, where $\epsilon_i(t)$ is white gaussian noise with fixed variance. Functions are additionally observed over a vector of discrete grid of timepoints $\boldsymbol{t}_i = \left\{t_{ij} : j \in \{1, \ldots, J_i\}\right\}$ which may vary across subjects. Let $Y_i(t_{ij})$ be the value of

$Y_i(\cdot)$ evaluated at t_{ij}, and $Y_i(\boldsymbol{t}_i)$ to be the $J_i \times 1$ vector of $Y_i(\cdot)$ evaluated over $\boldsymbol{t}_i$; similar notation will be used for other functions so that e.g. $\mu(\boldsymbol{t}_i)$ is the vector containing the mean $\mu(\cdot)$ evaluated over $\boldsymbol{t}_i$. We express the mean function $\mu(t)$ and eigenfunctions $\boldsymbol{\Phi}(t)$ using a spline basis $\mathbf{W}(t) = \{w_p(t) : p \in \{1, \ldots, P\}\}$. Let $\mathbf{W}(t_{ij})$ be a $1 \times P$ vector of the spline basis evaluated at t_{ij} and $\mathbf{W}(\boldsymbol{t}_i)$ be the $J_i \times P$ matrix of spline basis evaluated over the vector $\boldsymbol{t}_i$; $\boldsymbol{\beta}_\mu$ be a $P \times 1$ vector of spline coefficients corresponding to $\mu(t)$; and $\boldsymbol{\beta}_{\boldsymbol{\Phi}} = \left[\boldsymbol{\beta}_{\phi_1}, \ldots, \boldsymbol{\beta}_{\phi_K}\right]$ be the $P \times K$ matrix of spline coefficients corresponding to $\boldsymbol{\Phi}(t)$. Thus, we have that $\mu(\boldsymbol{t}_i) = \mathbf{W}(\boldsymbol{t}_i)\boldsymbol{\beta}_\mu$ is a $J_i \times 1$ vector and $\boldsymbol{\Phi}(\boldsymbol{t}_i) = [\phi_1(\boldsymbol{t}_i), ..., \phi_K(\boldsymbol{t}_i)]^T = \mathbf{W}(\boldsymbol{t}_i)\boldsymbol{\beta}_{\boldsymbol{\Phi}}$ is a $J_i \times K$ matrix. Lastly, let $\boldsymbol{\xi}_i = \{\xi_{ik} : k \in \{1, \ldots, K\}\}$ be a $K \times 1$ vector of observation-specific scores.

With the preceding notation, the working model of the observed data $Y_i(\cdot)$ measured at timepoint t_{ij} is given by

$$\begin{aligned} Y_i(t_{ij}) &= \mu(t_{ij}) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t_{ij}) + \epsilon_i(t_{ij}) \\ &= \mathbf{W}(t_{ij})\boldsymbol{\beta}_\mu + \mathbf{W}(t_{ij})\boldsymbol{\beta}_{\boldsymbol{\Phi}}\xi_i + \epsilon_i(t_{ij}) \end{aligned} \tag{2.2}$$

where $\epsilon_i(t_{ij})$ is additional noise with unknown var $\left[\epsilon_i(t_{ij})\right] = \sigma_\epsilon^2$. Under the common distributional assumptions that $\epsilon_i(t_{ij}) \sim \mathrm{N}\left(0, \sigma_\epsilon^2\right)$ and $\boldsymbol{\xi}_i \sim \mathrm{MVN}\left(0, \mathbf{I}_{K\times K}\right)$ [15], the model (2.2) is estimated by finding the maximum likelihood estimates (MLE) of the spline coefficients $\boldsymbol{\beta}_\mu$ and $\boldsymbol{\beta}_{\boldsymbol{\Phi}}$, the scores $\boldsymbol{\Xi} = \left\{\boldsymbol{\xi}_i : i \in \{1, \ldots, I\}\right\}$, and the variance of the error σ_ϵ^2.

2.3.2 Adaptive Scatterplot Smoothing via an Adaptive Ridge Penalty

The novel approach to adaptive scatterplot smoothing presented in this subsection is a central contribution of this manuscript, and later will be used in the context of FPCA. To the extent possible, we retain notation introduced in the previous section. We observe data $\{(t_i, y_i) : i \in \{1, \ldots, I\}\}$ and assume $y_i = f(t_i) + \epsilon_i$ where $\epsilon_i \sim N\left(0, \sigma_\epsilon^2\right)$. The goal of scatterplot smoothing is to flexibly estimate the unknown function $f(\cdot)$ defined over $t \in \mathcal{T}$. An spline-based estimator of $f(t)$ can be

obtained by expanding $f(t_i) = \mathbf{W}(t_i)\boldsymbol{\beta}_f$ using the spline basis $\mathbf{W}(t)$ and the vector of coefficients $\boldsymbol{\beta}_f$, and maximizing the likelihood or, equivalently, minimizing a residual sum of squares with respect to $\boldsymbol{\beta}_f$.

Spline-based methods can explicitly protect against overfitting by imposing a penalty on the complexity of the estimate $\hat{f}(t)$; squared-second-derivative penalties are a common choice. Define $\mathbf{W}''(t) = \left\{w_p''(t) : p \in \{1, \ldots, P\}\right\}$ to be the second derivatives of the spline functions $\mathbf{W}(t)$. Then $\lambda \int_{\mathcal{T}} [f''(t)]^2 \, dt = \lambda \sum_{p=1}^{P} \sum_{q=1}^{P} \beta_{fp}\beta_{fq} \int_{\mathcal{T}} w_p''(t) w_q''(t) dt$ is the squared-second derivative penalty, and λ is the associated tuning parameter. More compactly, define $\boldsymbol{\Omega}$ to be the $P \times P$ penalty matrix with entries

$$\boldsymbol{\Omega}_{pq} = \int_{\mathcal{T}} w_p''(t) w_q''(t) dt, \quad p, q \in \{1, \ldots, P\} \tag{2.3}$$

so that $\lambda \int_{\mathcal{T}} [f''(t)]^2 \, dt = \lambda \boldsymbol{\beta}_f^T \boldsymbol{\Omega} \boldsymbol{\beta}_f$. The integrals in (2.3) are known or can be estimated numerically. Define $\mathbf{y} = \{y_i : i \in \{1, \ldots, I\}\}$ and $\boldsymbol{t} = \{t_i : i \in \{1, \ldots, I\}\}$ to be $1 \times I$ observation vectors and $\mathbf{W}(\boldsymbol{t})$ to be a $I \times P$ matrix containing values of $\mathbf{W}(t)$ evaluated over $\boldsymbol{t}$. Then the penalized likelihood used to estimate $f(\cdot)$ is

$$\ell(\boldsymbol{\beta}_f, \sigma_\epsilon^2, \boldsymbol{\Omega}; \boldsymbol{y}, \boldsymbol{t}) = -\frac{I}{2} log(\sigma_\epsilon^2) - \frac{I}{2\sigma_\epsilon^2} \left\| \mathbf{y} - \mathbf{W}(\boldsymbol{t})\boldsymbol{\beta}_f \right\|^2 + \lambda \boldsymbol{\beta}_f^T \boldsymbol{\Omega} \boldsymbol{\beta}_f. \tag{2.4}$$

A mathematically equivalent likelihood can be obtained by treating spline coefficients $\boldsymbol{\beta}$ as random effects with a covariance proportional to $\boldsymbol{\Omega}^{-1}$. Doing so relates the tuning parameter to the variance of the random effects, so that the tuning parameter can be estimated from data rather than using a computationally expensive cross validation procedure. This observation underlines several approaches to scatterplot smoothing [23].

An adaptive smoothness penalty replaces the scalar tuning parameter λ by a tuning function $\lambda(t)$ defined over the domain $\mathcal{T}$. Our strategy for adaptive scatterplot smoothing is to connect this tuning function with an adaptive ridge penalty. We let $\lambda(t) = \left[\sum_{p=1}^{P} \lambda_p m_p(t)\right]^2$ where $\mathbf{M}(t) = \left\{m_p(t) : p \in \{1, \ldots, P\}\right\}$ is a spline expansion of the same dimension as $\mathbf{W}(t)$ and $\boldsymbol{\lambda} = \left\{\lambda_p : p \in \{1, \ldots, P\}\right\}$ is the corresponding $P \times 1$ vector of coefficients. The quadratic form

ensures the required constrain that $\lambda(t) \geq 0$ for all $t \in \mathcal{T}$. Thus, the adaptive smoothing penalty represented in terms of two sets of spline basis and respective coefficients is

$$\int_{\mathcal{T}} \lambda(t) f''(t)^2 dt = \sum_{p=1}^{P} \sum_{q=1}^{P} \sum_{r=1}^{P} \sum_{s=1}^{P} \left[\lambda_p \lambda_q \beta_{f_r} \beta_{f_s} \int_{\mathcal{T}} m_p(t) m_q(t) w_r''(t) w_s''(t) dt \right]. \tag{2.5}$$

While this expression holds for any choice of $\mathbf{W}(t)$ and $\mathbf{M}(t)$, we will construct $\mathbf{W}^*(t)$ and $\mathbf{M}(t)$ from a starting $\mathbf{W}(t)$ such that the conditions

$$\begin{aligned} &\int_{\mathcal{T}} m_p(t) m_q(t) w_r^{*\prime\prime}(t) w_s^{*\prime\prime}(t) dt = 1 \text{ when } p = q = r = s \\ &\int_{\mathcal{T}} m_p(t) m_q(t) w_r^{*\prime\prime}(t) w_s^{*\prime\prime}(t) dt = 0 \text{ otherwise} \end{aligned} . \tag{2.6}$$

hold. That is, for choices of $\mathbf{W}^*(t)$ and $\mathbf{M}(t)$ such that 2.6 holds,

$$\int_{\mathcal{T}} \lambda(t) f''(t)^2 dt = \sum_{p=1}^{P} \lambda_p^2 \boldsymbol{\beta}_{f_p}^2 = \boldsymbol{\beta}_f^T \boldsymbol{\Lambda} \boldsymbol{\beta}_f \tag{2.7}$$

where $\boldsymbol{\Lambda} = \text{diag}\left\{\lambda_1^2, \ldots, \lambda_p^2\right\}$ is a $P \times P$ diagonal matrix. In this case, the adaptive smoothing penalty (2.5) becomes an adaptive ridge penalty in which the spline coefficients $\boldsymbol{\beta}_f$ for the function $f(\cdot)$ are weighted by the spline coefficients $\boldsymbol{\lambda}$ for $\lambda(\cdot)$. The procedure to construct $\mathbf{W}^*(t)$ and $\mathbf{M}(t)$ such that (2.7) holds is covered later in this section; first we describe our estimation approach assuming these bases are available.

Define $\mathbf{W}^*(\boldsymbol{t})$ to be $I \times P$ matrix of values of $\mathbf{W}^*(t)$ evaluated at $\boldsymbol{t}$ and let $\boldsymbol{\beta}_f^*$ to be the corresponding coefficients for $\mathbf{W}^*(\boldsymbol{t})$, such that $\mathbf{W}^*(\boldsymbol{t})\boldsymbol{\beta}_f^* = \mathbf{W}(\boldsymbol{t})\boldsymbol{\beta}_f$. Then, the penalized likelihood used to estimate $f(\cdot)$ is

$$\ell(\boldsymbol{\beta}_f^*, \sigma_\epsilon^2, \boldsymbol{\Lambda}; \boldsymbol{y}, \boldsymbol{t}) = -\frac{I}{2} log(\sigma_\epsilon^2) - \frac{I}{2\sigma_\epsilon^2} \left\| \mathbf{y} - \mathbf{W}^*(\boldsymbol{t})\boldsymbol{\beta}_f^* \right\|^2 + \lambda \boldsymbol{\beta}_f^{*T} \boldsymbol{\Lambda}^* \boldsymbol{\beta}_f^* \tag{2.8}$$

where $\boldsymbol{\Lambda}^*$ is the corresponding penalty matrix for coefficients $\boldsymbol{\beta}_f^*$. Although the likelihood in

(2.8) resembles that in (2.4), the changes in the basis and penalty provide an approach to adaptive scatterplot smoothing.

We maximize (2.8) with an algorithm that iterates between updating coefficients $\boldsymbol{\beta}_f^*$, tuning parameters $\boldsymbol{\Lambda}^*$, and the residual variance σ_ϵ^2 using the following estimators

- $\hat{\boldsymbol{\beta}}_f^* | \boldsymbol{\Lambda}^*, \sigma_\epsilon^2 = \frac{1}{\sigma_\epsilon^2} \left[\frac{\mathbf{W}^*(\boldsymbol{t})^T \mathbf{W}^*(\boldsymbol{t})}{\sigma_\epsilon^2} + \boldsymbol{\Lambda}^* \right]^{-1} \mathbf{W}^*(\boldsymbol{t})^T \mathbf{y}$
- $\hat{\boldsymbol{\Lambda}}^* | \boldsymbol{\beta}_f^* = \text{diag}\left\{0, 0, \hat{\lambda}_3^2, \ldots, \hat{\lambda}_P^2\right\} \quad \text{where} \quad \hat{\lambda}_p = \frac{1}{\boldsymbol{\beta}_{f_p}^*}$
- $\hat{\sigma}_\epsilon^2 | \boldsymbol{\beta}_f^* = \frac{\|\mathbf{y} - \mathbf{W}^*(\boldsymbol{t}) \boldsymbol{\beta}_f^*\|^2}{I}$

we initialize our algorithm by letting $\boldsymbol{\Lambda}^{*(1)} = \mathbf{0}_{P \times P}$ and obtain unpenalized coefficients $\hat{\boldsymbol{\beta}}_f^{*(1)} = \left[\tilde{\mathbf{W}}(\boldsymbol{t})^T \mathbf{W}^*(\boldsymbol{t}) \right]^{-1} \mathbf{W}^*(\boldsymbol{t})^T \mathbf{y}$. At each iteration, we evaluate (2.8) given the current parameter estimates, and monitor the absolute difference between the current evaluated penalized likelihood and its previous estimate.

We now discuss our strategy for constructing bases $\mathbf{W}^*(t)$ and $\mathbf{M}(t)$ that satisfy the constraints in (2.6). Our approach follows [39], who used a similar transformation to obtain a simple mixed effects model representation for non-adaptive smoothing. We find a transformation $\mathbf{U}$ such that $\mathbf{W}^*(\boldsymbol{t}) = \mathbf{W}(\boldsymbol{t})\mathbf{U}$ by eigendecomposing $\boldsymbol{\Omega}$ into an orthogonal matrix $\boldsymbol{Q}$ containing the eigenvectors of $\boldsymbol{\Omega}$ and a diagonal matrix $\boldsymbol{\Psi}$ whose entries are the eigenvalues such that $\boldsymbol{\Omega} = \boldsymbol{Q}^T \boldsymbol{\Psi} \boldsymbol{Q}$. $\mathbf{W}(\boldsymbol{t})$ spans the space of straight lines, but the penalty is on second derivatives, thus $\text{rank}(\boldsymbol{\Omega}) = P - 2$, and $\boldsymbol{\Psi}$ will have two zero-entries and $P - 2$ positive entries [40]. Define partitions $\boldsymbol{Q} = [\boldsymbol{Q}_1 | \boldsymbol{Q}_2]$ and $\Psi = [\boldsymbol{\Psi}_1 | \boldsymbol{\Psi}_2]$ such that $\boldsymbol{Q}_2$ and $\boldsymbol{\Psi}_2$ are sub-matrices of $\boldsymbol{Q}$ and $\boldsymbol{\Psi}$ with columns that correspond to the non-zero eigenvalues in $\boldsymbol{\Psi}$, then $\mathbf{U} = \left[\mathbf{Q}_1 | \boldsymbol{Q}_2 \boldsymbol{\Psi}_2^{-1/2} \right]$. For this choice of $\mathbf{U}$, we have that

$$\mathbf{W}^{*\prime\prime}(\boldsymbol{t})^T \mathbf{W}^{*\prime\prime}(\boldsymbol{t}) = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{(P-2)\times(P-2)} \end{bmatrix} \tag{2.9}$$

so $\mathbf{M}(\boldsymbol{t}) = \left\{ m_q(\boldsymbol{t}) = \frac{\boldsymbol{\beta}_{f_q}^* w_q^{*\prime\prime}(\boldsymbol{t})}{\sum_{p=3}^P \boldsymbol{\beta}_{f_p}^* w_p^{*\prime\prime}(\boldsymbol{t})} : q \in \{3, \ldots, P\} \right\}$ will satisfy the conditions in (2.6); as a result adaptive scatterplot smoothing can indeed be implemented via an adaptive ridge penalty (2.7).

We note that the basis $\mathbf{M}(t)$ does not affect the estimation of $\boldsymbol{\beta}_f^*$ or $\boldsymbol{\Lambda}^*$; it is sufficient to know that such a basis exists for the algorithm to be well-defined. Since $\mathbf{M}(t)$ depends on the current estimate of $\boldsymbol{\beta}_f^*$, the basis varies across iterations (although $W^*(t)$ remains fixed). At any iteration, however, the current estimate of the penalty function can be obtained as $\hat{\lambda}(t) = \left[\frac{\sum_{p=1}^{P} \hat{\lambda}_p \hat{\beta}_{fp} w_p^{*\prime\prime}(t)}{\sum_{p=1}^{P} \hat{\beta}_{fp} w_p^{*\prime\prime}(t)}\right]^2$. Under this framework, each of the penalized coefficients and their corresponding spline basis carries different local and global information about the overall smoothness of the estimated fit, and allowing each of these splines to be weighted differently relaxes the assumption that the smoothness across the fit is equal.

Our work draws parallelism to previous work on smoothing splines and mixed effects models. Our estimators are the equivalent solution to a random effects model with each of the coefficients having a separate Gaussian prior with unique variance. In our implementation, the elements of $\boldsymbol{\beta}_f^*$ to be greater than 10^{-6}, as the elements of $\boldsymbol{\beta}_f^*$ can converge towards very small non-zero values, leading to extremely large values in the diagonal of $\boldsymbol{\Lambda}^*$, and iteration can make that to be numerically infeasible.

2.3.3 Adaptive Smoothing Functional Principal Component Analysis

Our primary objective in this manuscript is to estimate patterns of variation shared across functional observations using FPCA; the latent functions are also used to reconstruct and denoise individual curves. We smooth each FPCA adaptively to capture local variations in smoothness across the functional domain. We extend the FPCA framework in Section 2.3.1 to estimate the mean function $\mu(t)$ and set of FPCs $\boldsymbol{\Phi}(t) = \{\phi_k(t) : k \in \{1, \ldots, K\}\}$ using the the adaptive smoothness penalty developed for scatterplot smoothing in Section 2.3.2. We now describe our methodology in detail.

Express $\mu(t_i) = \mathbf{W}^*(\boldsymbol{t}_i)\boldsymbol{\beta}_\mu^*$ and $\boldsymbol{\Phi}(\boldsymbol{t}_i) = [\phi_1(\boldsymbol{t}_i), ..., \phi_K(\boldsymbol{t}_i)]^T = \mathbf{W}^*(\boldsymbol{t}_i)\boldsymbol{\beta}_{\boldsymbol{\Phi}}^*$ using the transformed cubic B-spline discussed in Section 2.3.2 and corresponding coefficients $\boldsymbol{\beta}_\mu^*$ and $\boldsymbol{\beta}_{\boldsymbol{\Phi}}^*$. We maximize:

$$
\begin{aligned}
&\ell\left(\boldsymbol{\beta}^*_\mu, \boldsymbol{\beta}^*_{\boldsymbol{\Phi}}, \sigma^2_\epsilon; \boldsymbol{Y}, \boldsymbol{\Xi}\right) - P(\boldsymbol{\beta}^*_\mu, \boldsymbol{\beta}^*_{\boldsymbol{\Phi}}) = \\
&\ell\left(\boldsymbol{\beta}^*_\mu, \boldsymbol{\beta}^*_{\boldsymbol{\Phi}}, \sigma^2_\epsilon; \boldsymbol{Y}|\boldsymbol{\Xi}\right) + \ell(\boldsymbol{\Xi}) - P(\boldsymbol{\beta}^*_\mu, \boldsymbol{\beta}^*_{\boldsymbol{\Phi}}) \sim \\
&\sum_{i=1}^{I}\left(\frac{\|Y_i(\boldsymbol{t}_i) - \mathbf{W}^*(\boldsymbol{t}_i)\boldsymbol{\beta}^*_\mu - \mathbf{W}^*(\boldsymbol{t}_i)\tilde{\boldsymbol{\beta}}_{\boldsymbol{\Phi}}\boldsymbol{\xi}_i\|^2}{2\sigma^2_\epsilon}\right) + \sum_{i=1}^{I}\frac{\left\|\boldsymbol{\xi}_i\right\|^2}{2} + \sum_{i=1}^{I}\frac{J_i \log\sigma^2_\epsilon}{2} - P(\boldsymbol{\beta}^*_\mu, \boldsymbol{\beta}^*_{\boldsymbol{\Phi}})
\end{aligned}
\tag{2.10}
$$

with respect to $\boldsymbol{\beta}^*_\mu$ and $\boldsymbol{\beta}^*_{\boldsymbol{\Phi}}$, σ^2_ϵ and $\boldsymbol{\Xi}$. The penalty term $P(\boldsymbol{\beta}^*_\mu, \boldsymbol{\beta}^*_{\boldsymbol{\Phi}})$ is defined as:

$$
P(\boldsymbol{\beta}^*_\mu, \boldsymbol{\beta}^*_{\boldsymbol{\Phi}}) = \int_{\mathcal{T}} \lambda_\mu(t)\mu''(t)^2 dt + \sum_{k=1}^{K}\left(\int_{\mathcal{T}} \lambda_{\phi_k}(t)\phi''_k(t)^2 dt\right). \tag{2.11}
$$

using the techniques described in Section 2.3.2 this penalty can be expressed as

$$
{\boldsymbol{\beta}^*_\mu}^T \boldsymbol{\Lambda}_\mu \boldsymbol{\beta}^*_\mu + \sum_{k=1}^{K} {\boldsymbol{\beta}^*_{\phi_k}}^T \boldsymbol{\Lambda}_{\phi_k} \boldsymbol{\beta}^*_{\phi_k} \tag{2.12}
$$

where $\boldsymbol{\Lambda}_\mu$ and $\boldsymbol{\Lambda}_{\phi_k}$, $k = 1, \ldots, K$, are diagonal $P \times P$ matrices of corresponding tuning parameters (e.g. $\boldsymbol{\Lambda}_\mu = \text{diag}\left[0, 0, \lambda_{\mu 3}, \ldots, \lambda_{\mu P}\right]$). As in Section 3.2, this provides direct estimates of the tuning functions $\lambda_\mu(t)$ and $\lambda_{\phi_k}(t)$, $k = 1, \ldots, K$, as part of the adaptive FPCA approach.

Algorithm and Implementation

We now proceed to describe our iterative approach to estimate our model. Our algorithm iterates between the following steps:

1. Subject-specific scores $\boldsymbol{\xi}_i$ are estimated given current estimates of coefficients $\boldsymbol{\beta}^*_\mu$ and $\boldsymbol{\beta}^*_{\boldsymbol{\Phi}}$, and residual variance σ^2_ϵ.

2. $\mu(t)$ and $\boldsymbol{\Phi}(t)$ are estimated by maximizing (2.10) with respect to $\boldsymbol{\beta}^*_\mu$ and $\boldsymbol{\beta}^*_{\boldsymbol{\Phi}}$ conditional on the current estimates of scores $\boldsymbol{\Xi}$, tuning parameters, and σ^2_ϵ.

3. Parameters $\boldsymbol{\Lambda}_\mu$ and $\boldsymbol{\Lambda}_{\phi_k}$, $k = 1, \ldots, K$ are tuned as described in Section 2.3.2 given current values of $\boldsymbol{\beta}^*_\mu$ and $\boldsymbol{\beta}^*_{\boldsymbol{\Phi}}$. The residual variance σ^2_ϵ is estimated using the method of moments.

The initialization of this algorithm is discussed below. At each iteration, we evaluate (2.10) given the current parameter estimates, and monitor the absolute difference between the current evaluated penalized likelihood and its previous estimate.

Functional Principal Component Scores Estimation

At any given iteration, the subject-specific scores are calculated by maximizing the full likelihood of (2.10) given all other parameters are known. For observation $Y_i(\boldsymbol{t}_i)$ the estimate of $\boldsymbol{\xi}_i$ is:

$$\hat{\boldsymbol{\xi}}_i|Y_i(\boldsymbol{t}_i), \mu(\boldsymbol{t}_i), \boldsymbol{\Phi}(\boldsymbol{t}_i), \sigma_\epsilon^2 = \left(\frac{\boldsymbol{\Phi}(\boldsymbol{t}_i)^T\boldsymbol{\Phi}(\boldsymbol{t}_i)}{\sigma_\epsilon^2} + \mathbf{I}_{K\times K}\right)^{-1} \boldsymbol{\Phi}(\boldsymbol{t}_i)^T(Y_i(\boldsymbol{t}_i) - \mu(\boldsymbol{t}_i)) \tag{2.13}$$

Latent Functions Estimation

Define $\boldsymbol{Y} = [Y_1(\boldsymbol{t}_1), ..., Y_I(\boldsymbol{t}_I)]^T$ be to be a $\sum_{i=1}^I (J_i) \times 1$ vector of concatenated vectors $Y_i(\boldsymbol{t}_i)$. For each subject i, denote $\boldsymbol{\theta}_i = (1, \boldsymbol{\xi}_i)$ to be a vector of dimension $1 \times (K+1)$ and define the matrix $\boldsymbol{\Theta}_i(\boldsymbol{t}_i) = \boldsymbol{\theta}_i \otimes \mathbf{W}^*(\boldsymbol{t}_i)$ where $\otimes$ is the Kronecker product. Further, let $\boldsymbol{B^*} = [\boldsymbol{\beta}_\mu^{*T}, \boldsymbol{\beta}_{\phi_1}^{*T}, \ldots, \boldsymbol{\beta}_{\phi_K}^{*T}]^T$ to be a $(K+1)P \times 1$ vector of concatenated coefficient vectors. Using this notation we have that (2.2) can be rewritten as $\mathbf{Y} = \boldsymbol{\Theta B^*} + \boldsymbol{E}$, where $\boldsymbol{E}$ is the $\sum_{i=1}^I J_i \times 1$ vectorized version of the error term ϵ_{ij} and $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1(\boldsymbol{t}_1)^T, \ldots, \boldsymbol{\Theta}_I(\boldsymbol{t}_I)^T]^T$. Define $\boldsymbol{\Lambda}_\mu = \text{diag}(\lambda_{\mu 1}, \ldots, \lambda_{\mu p})$ and $\boldsymbol{\Lambda}_{\phi_k} = \text{diag}(\lambda_{\phi_k 1}, \ldots, \lambda_{\phi_k p})$ for each $k \in 1, \ldots, K$ to be a $P \times P$ diagonal matrix of tuning parameters, and $\mathbf{0}_{P\times P}$ to be the zero matrix, then we have that the solution for $\boldsymbol{B^*}$ is simply:

$$\hat{\boldsymbol{B}}^*|\mathbf{Y}, \boldsymbol{\Theta}, \sigma_\epsilon^2, \boldsymbol{\Lambda} = \frac{1}{\sigma_\epsilon^2}\left(\frac{\boldsymbol{\Theta}^T\boldsymbol{\Theta}}{\sigma_\epsilon^2} + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{\Theta}^T Y \tag{2.14}$$

where $\boldsymbol{\Lambda}$ is equal to

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{\mu} & \mathbf{0}_{P\times P} & \dots & \mathbf{0}_{P\times P} \\ \mathbf{0}_{P\times P} & \mathbf{\Lambda}_{\phi_1} & \dots & \mathbf{0}_{P\times P} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{P\times P} & \mathbf{0}_{P\times P} & \dots & \mathbf{\Lambda}_{\phi_K} \end{pmatrix}. \tag{2.15}$$

Tuning Adaptive Smoothness Penalty

At each iteration, we update the adaptive smoothing weights given the present estimates of the coefficients, so that at the next iteration the coefficients are estimated with an updated penalty matrix. We do this in the same fashion as our adaptive scatterplot algorithm 2.3.2. Given $\boldsymbol{B}^*$ we have $\mathbf{\Lambda}_{\mu} = \text{diag}\left(0, 0, (\frac{1}{\tilde{\beta}_{\mu 3}})^2, \dots, (\frac{1}{\tilde{\beta}_{\mu P}})^2\right)$ and $\mathbf{\Lambda}_{\phi_k} = \text{diag}\left(0, 0, (\frac{1}{\tilde{\beta}_{\phi_k 3}})^2, \dots, (\frac{1}{\tilde{\beta}_{\phi_k P}})^2\right)$ for $k \in \{1, \dots, K\}$ as the updated estimates of the respective tuning parameters for each set of coefficients. The first two entries of $\mathbf{\Lambda}_{\mu} = \text{diag}(\lambda_{\mu 1}, \dots, \lambda_{\mu p})$ and $\mathbf{\Lambda}_{\phi_k} = \text{diag}(\lambda_{\phi_k 1}, \dots, \lambda_{\phi_k p})$ for each $k \in 1, \dots, K$ are zero as in 2.3.2. The residual variance σ^2_{ϵ} is estimated via method of moments.

Practical Concerns

The algorithm can be initialized using random values for $\mathbf{\Xi}^{(0)}$ where $\boldsymbol{\xi}_i^{(0)} \sim N(0, \mathbf{I}_{K\times K})$. However, our default setting is to start the algorithm by using the scores from `refund::fpca.face`, as this leads to faster convergence. After initializing $\mathbf{\Xi}$, we set tuning parameters as $\mathbf{\Lambda}_{\mu}^{(1)} = \mathbf{\Lambda}_{\phi_k}^{(1)} = \mathbf{0}_{P\times P} \quad \forall k \in 1, \dots, K$, and find initial estimates of $\boldsymbol{\beta}_{\mu}^*$, $\boldsymbol{\beta}_{\Phi}^*$, and σ^2_{ϵ} using ordinary least squares.

We briefly note some practical concerns related to probabilistic approaches to PCA and FPCA, including our adaptive FPCA method. First, like many probabilistic approaches to PCA and FPCA, we do not constrain the FPCs to be orthogonal as part of the estimation algorithm. Orthogonality can be enforced after the algorithm converges; we have observed faster convergence when orthogonalizing FPCs in each iteration, and take this approach in our implementation. Second, in contrast to eigendecomposition methods, we need to pre-specify the number of FPCs to be estimated in our model. It is common to threshold the number of FPCs by only keeping the most informative

ones using the percent-of-variance explained. We address this by defaulting to modeling a large number of FPCs, as in most settings, this will capture the majority of the variability in the data. Our implementation defaults to modeling 15 FPCs, and defaults only to keep the most informative FPCs that explain 99% of the variability in the data.

2.4 Simulations

We demonstrate the performance of our proposed method using simulated data to mimic real-world examples, such as our motivating study. We compare our method to the standard FPCA decomposition used in FDA. We examine the ability of our Adaptive Smoothing FPCA to 1) Recover functional principal components, 2) Reconstruct curves based on our estimated FPC expansion, and 3) Compare computational efficiency by run time.

2.4.1 Simulation Design

We generate curves $Y_i(t)$ in which the level of smoothness varies temporally according to the FPCA model:

$$Y_i(t) = \mu(t) + \sum_{k=1}^{2} \xi_{ik}\phi_k(t) + \epsilon_i(t) \tag{2.16}$$

We let $t \in [0, 1]$ to be an equally spaced grid of fixed length that is shared across all observations. This represent the period of time in which neuron measurements of activation are undertaken in the context of our motivating example. We generate data from piece-wise functions, in which the first half of the grid is constant, and the second half of the grid has high variation over time. Thus, incorporating varying temporal smoothness. We generate data from a constant mean $\mu(t) = t^{\frac{-3}{2}} sin(\pi t^{\frac{1}{4}}) * I(t > \frac{1}{2})$ and functional principal components with varying temporal smoothness by letting $\phi_k(t) = t^{\frac{-3}{2}} sin(4k\pi t^{\frac{1}{4}}) * I(t > \frac{1}{2})$. This FPCs are sine functions with temporally varying periods, they have high variation over time, and are orthogonal to each other. The scores ξ_{ik} are generated from a normal distribution with $\sigma_k^2 = 4, 1$ for $k = 1, 2$ respectively. Furthermore

we let add white noise to our observations by letting $\epsilon_i(t_{ij}) \sim N(0, \sigma_\epsilon^2)$ for fixed values of σ_ϵ^2.

We ran our simulations using five cores on an M1 Pro processor. We evaluate the algorithm's performance as a function of the sample size and noise level. We simulate observations with 100 timepoints each, and generate 100 datasets for each combination of sample size (25, 50, 100), and error variance (0.1, 0.2). We apply the method described in Section 3 for each simulated dataset, and compare with `refund::fpca.sc` [9, 41], a standard implementation of FPCA, and `refund::fpca.face` [14, 41] which is optimized for speed. We fix the number of spline basis to be 40 across all three implementations and allow the three implementations to pick the final number of FPCs based on 99% percent variance explained (PVE). All other settings in these methods are left at default values.

We compare standard FPCA to our method in estimation accuracy and computation time. We quantify accuracy by calculating the integrated square error (ISE). We first estimate the ISE for the true mean and functional principal components. We average the ISE for each component within the same simulation setting. The mean ISE (MISE) quantifies how well each method can reconstruct the latent functions in our data. We further estimate the observation-specific ISE and calculate the MISE across all observations within the same simulation setting to estimate the accuracy of the curve reconstruction.

2.4.2 Simulation Results

Figure 2.2 summarizes the fit of simulated datasets for one representative simulation setting (sample size = 25, noise variance = 0.2). Panel A shows the estimates of the true data-generating curves. From left to right, the columns show the estimates across the mean $\mu(t)$, and each of the functional principal components $\phi_1(t), \phi_2(t)$. Each black line shows an estimate from each simulated dataset. When the data-generating mechanism exhibits sharp changes in smoothness over time, our method is able to capture these complex changes in local smoothness, while the conventional approaches fail to properly capture these sharp changes, as displayed in the bottom two rows. These approaches tend to under-smooth in low-smoothness sections in order to properly

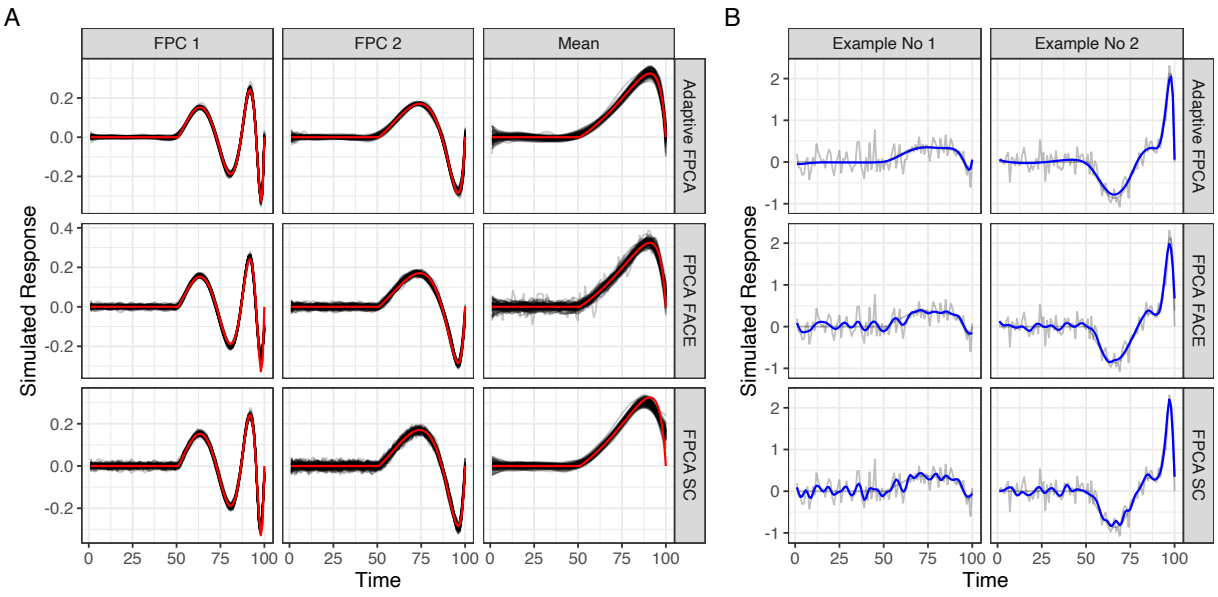


Figure 2.2: This figure shows the estimates of the the true data-generating functions, and sample observations as estimated by our proposed method and the standard approaches (`fpca.sc` and `fpca.face`). This plot display a representative simulation scenario when $n = 25$, and $\sigma^2 = 0.2$. Panel A overlays the true data generating curves (in red) over the estimated data-general curves across 100 simulations. Panel B shows two representative curves, and the corresponding reconstructions estimated by each of the methods.

capture the sharp changes in smoothness across the entire curve. Panel B shows two representative examples of simulated-curves. This "trade-off" that conventional approaches introduce is evident in the under-smoothing the reconstructed curves have in the bottom two rows. In contrast, our approach in the top row is able to properly reconstruct the curve without introducing any artifacts.

Figure 2.3 summarizes the accuracy results across simulated datasets at different sample sizes and noise levels. Panel A summarizes the accuracy of estimating the true data-generating piece-wise functions. From left to right, the columns show the MISE across the mean $\mu(t)$, and each of the functional principal components $\phi_1(t), \phi_2(t)$. Panel B summarizes the accuracy in reconstructing individual curves across these simulation settings. For both, accuracy in estimating the data-generating functions, and curve reconstruction adaptive FPCA outperforms `fpca.sc` and `fpca.face` in terms of MISE. These results are consistent with the observations from Figure 2.2.

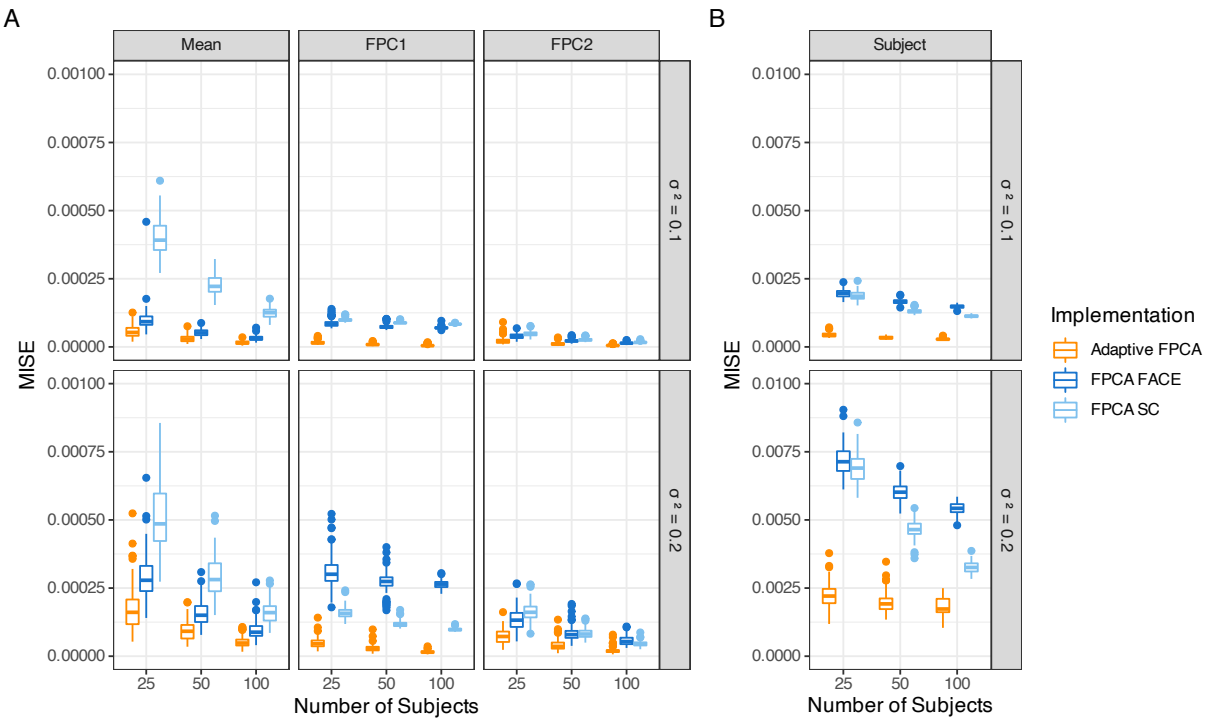


Figure 2.3: This plot shows the mean integrated square error (MISE) across the six simulation settings explored in our design. Panel A shows the accuracy of our proposed method (in orange), and the two standard approaches (in blue) when it comes to estimating the true data-generating functions. Panel B is a plot of MISE curve-specific reconstruction.

Figure 2.4 summarizes additional performance metrics of our method. Panel A shows the run time of each of the methods compared across all simulation scenarios. From this plot, we observe that our method is faster than `fpca.sc`. However, it is not able to outperform `fpca.face`, which is expected given that this implementation was optimized for speed. Panel B shows the median number of FPCs selected across each of the simulation settings. Our method better captures the true number of FPCs in the generating dataset compared with both `fpca.sc` and `fpca.face`. This is particularly relevant for high-noise settings, as either of those implementations selects a larger number of FPCs to explain the variability in the data, leading to more FPCs that are harder to interpret.

2.5 Application Results

The goal of our analysis is to derive patterns of activation that capture the underlying biological behavior of the motor cortex, as these will help us understand the dynamics between brain activation and its involvement in skilled movements [42]. Our dataset contains trial-averaged spike data for 25 neurons across the duration of each experimental trial (1.75 seconds) aligned to the auditory cue 2.1. We apply the Adaptive Functional Principal Component method described in Section 2.3 and Standard FPCA as implemented in `fpca.sc`. We fix the number of splines basis to be 40 and pre-specify the number of estimated FPCs from the dataset to be 15 in both implementations. We do a post hoc selection of an appropriate number of FPCs that explain 99% of the variability in the data as measured by the eigenvalues. Adaptive FPCA explains 99% of the variability in the data with 4 FPCs, with each capturing 87.76%, 6.73%, 3.30%, and 0.86%. Conventional FPCA explains 99% of the variability in the data with 5 FPCs, with each capturing 87.76%, 6.73%, 3.30%, 0.86%, and 0.56%.

Figure 5 shows the estimated mean, first two FPCs, and example curve reconstructions from adaptive FPCA and conventional FPC. Our method, displayed in the top row, better captures the biological processes behind neural activation during dexterous movement. Adaptive FPCA correctly captures the reaction period with no change in initial baseline activation while explaining the sharp change in neural activity after the reaction time and steady return to baseline over time. In contrast, standard FPCA introduces undersmoothing artifacts in the fit to properly account for the sharp changes in activation the neuron exhibits over the experiment. These results are consistent with our simulation analyses in Section 2.4.

The first two columns of graphs in Figure 5 (Activation Pattern 1 and Activation Pattern 2) show the activation patterns explained by the first two FPCs. These first two adaptive FPCs explain 95.14% of the variance, while the first two conventional FPCs explain 94.7%. The first FPC explains the variability in the change in activation after the reaction time and the amplitude of the initial activation peak. The second FPC explains an activation trade-off. Neurons with a high

increase in activation immediately post stimuli have lower baseline levels after the initial change. Conversely, neurons that experience decreased activation post stimuli return to a higher baseline level. The remaining three columns show distinct observed neural activation patterns and the reconstructions from both methods. These three patterns vary in the degree to which the neurons activate post-baseline and the degree to which they return to a baseline activation. Our adaptive FPCA reconstructions match the expected biological behavior of neurons in this experiment in contrast with the under-smoothed reconstructions from conventional FPCA.

2.6 Discussion

We present a novel approach to estimating a functional principal component decomposition that incorporates adaptive smoothness into estimating model components. We introduce a new implementation of adaptive scatterplot smoothing that is scalable to functional data analysis. We incorporate our proposed adaptive scatterplot smoothing technique into FPCA using a penalized likelihood framework. Our simulations suggest that our method is better suited to model data exhibiting sharp smoothness changes over time.

We are motivated by neural spike data from a mouse trained to reach for a pellet. We found activation patterns that explain the changes in the motor cortex during the experiment by applying our method to our motivating dataset. Future analysis will examine how these activation patterns derived from our data relate to dexterous movement. These results could be further used in downstream analysis to understand the relationship between brain activation and paw movement using a function-on-function regression approach.

While we assume that our data is generated from a Gaussian process, we know that the raw neural spikes could be modeled as binary or count data. Furthermore, we averaged across trials, as our exploratory analysis of the data suggests that the data-generating process is the same across all trials, and averaging across them leads to better estimates due to the noisiness of the raw data. Potential extensions of our method include generalizing Adaptive FPCA to non-gaussian distributions and multilevel approaches. Other extensions of interest could explore whether adaptive

smoothing can borrow information across curves to co-localize the estimated sharp changes.

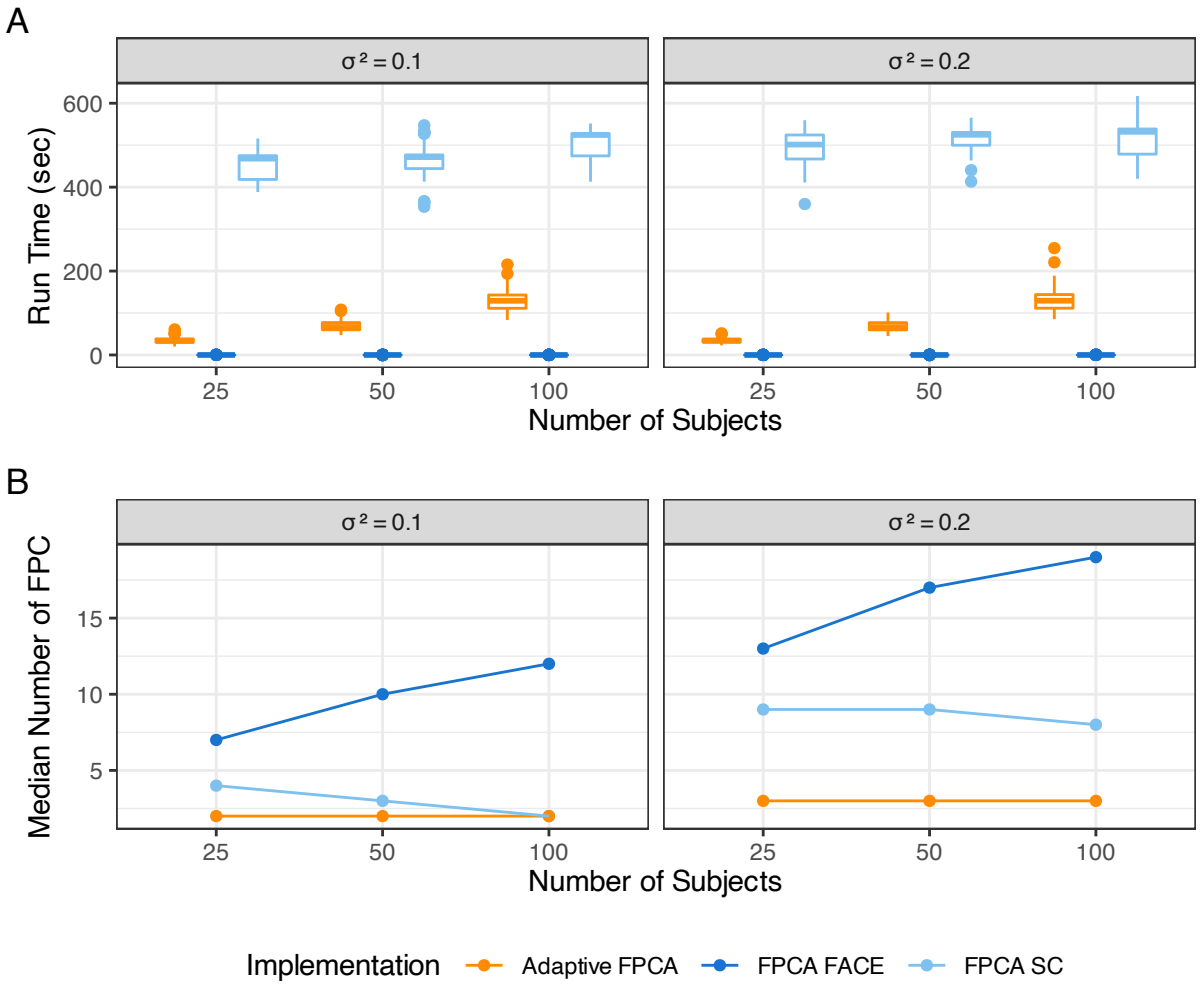


Figure 2.4: Panel A shows computation time of our proposed method (in orange), `fpca.sc` (in light blue) and `fpca.face` (in dark blue) across all six simulation settings. Panel B shows the median number of FPCs selected across the 100 simulation within each of the simulation settings.

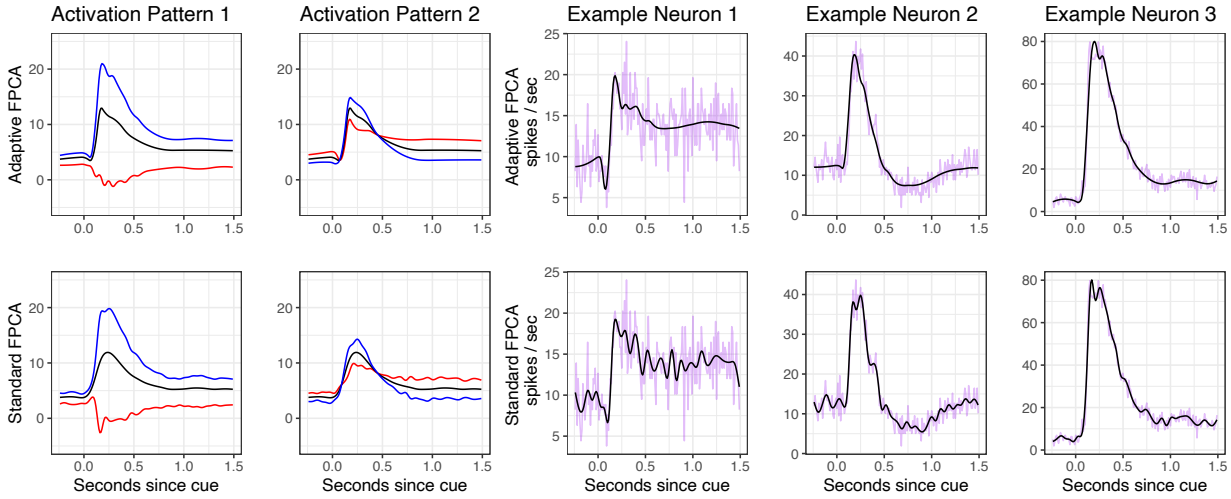


Figure 2.5: Results of applying our methodology and standard FPCA to our motivating dataset. The top row shows the estimates derived from Adaptive FPCA. The bottom row shows the same estimates for standard FPCA implemented using `fpca.sc`. The first two columns shows the variability explained by the first two activation patterns (these plots show the mean activation (in black) plus (in blue) activation pattern times 75^{th} score quantile and plus (in red) activation pattern times 25^{th} score quantile. The last three columns are the curve reconstructions for Neuron 2-4 in Figure 1.

Chapter 3: Two Sample Test for Eigendecompositions of Functional Data

3.1 Introduction

In-vivo cell-specific neural spike measurements from animals have been fundamental in understanding the relationship between the brain and behavior [43, 44, 45, 42]. Neuroscientists commonly assume neural-spikes to be realizations of underlying latent neural activation process [46]. These brain activation patterns better explain behavior when compared to only using neuron-level observations [47, 48]. A growing body of literature suggests brain patterns change in response to new experimental conditions [49, 50, 51], and that activation patterns remain constant with the same experimental conditions [52, 53]. Thus it is commonly assumed that when stimuli remain constant, such as when an animal performs the same task repeatedly, activation patterns can be best estimated by aggregating information across trials. Our work uses data from an experiment in which a trained mouse reached for a food pellet after hearing an auditory cue [42]. The scientific goal of this manuscript is to test whether activation patterns from the motor cortex hold constant when a mouse repeatedly performs a reaching movement. Novel statistical techniques to compare such patterns can both inform the derivation of such patterns from neural-spike data and help formally infer if any changes derived activation patterns are attributable to experimental variables or noise.

Our motivating data contains spike activity for 25 neurons in the motor cortex recorded using silicon probes connected to the mouse's brain across 157 trials. Our data contains neural firing recorded in 10ms windows. Figure 3.1 shows activation in six representative neurons in the mouse's brain. Neurons exhibit a range of activation intensities after the auditory cue (at around 0.2 seconds) and distinct behaviors in how they return to the baseline activation level throughout the experiment. The behavior of each neuron remains relatively constant across each trial.

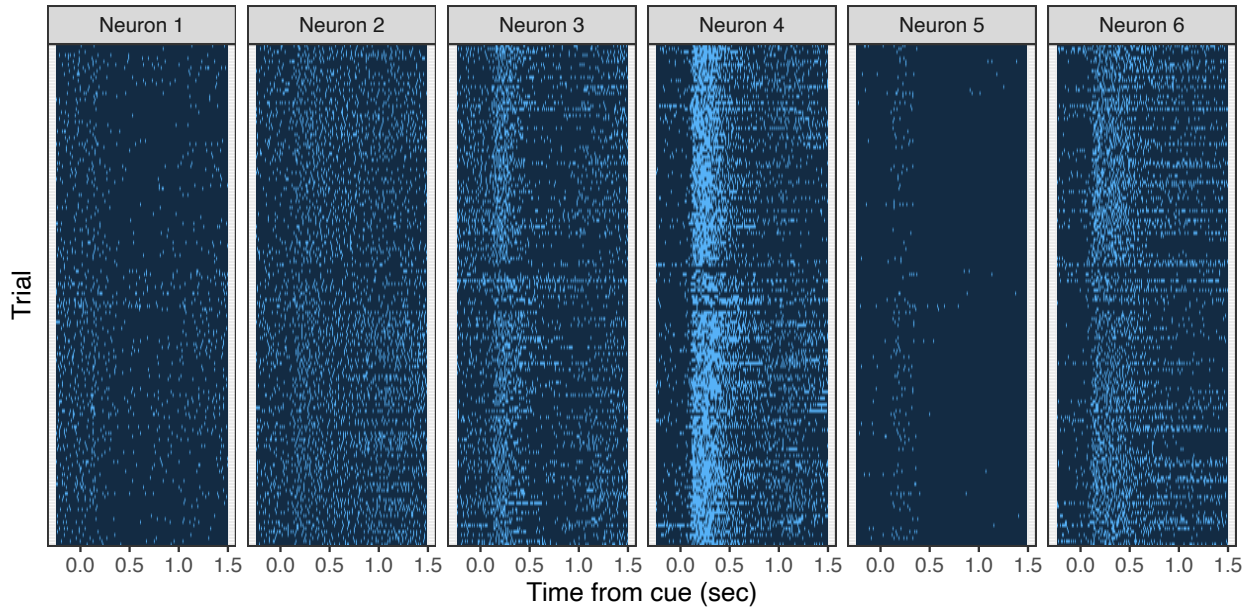


Figure 3.1: Panel A displays a lasagna plot of the activation of six example neurons across 174 timepoints and 157 trials. Light blue indicates that the neuron is activate at that specific instance.

Neural-spike data is assumed to come from an underlying data generating mechanism in which patterns give rise to the activation in the neuron level. Thus, we derive activation patterns from each trial using Functional Principal Component Analysis (FPCA) [9, 10]. We aim to test the broad scientific hypothesis that activation patterns obtained from an FPCA are the same. We first develop a test to test the equality of activation patterns in to trials by comparing score covariances; this is also a test of the covariance operators. Next, we use our proposed method to compare all pairs of trials and summarize the tests. We then generate a permutation-based null distribution of the summary statistic to test the global hypothesis that patterns are the same across all trials.

There are a number of statistical tests previously developed for two samples of functional data. Several tests focus on comparing the center of two groups, including the point-wise t-test to detect differences in sample functional means proposed by [1], L^2 norm-based tests [54, 55, 56, 54, 57, 58] or the likelihood-ratio approach proposed by [59]. Approaches to compare the distribution of two functional samples include proposed work by [56] to jointly test the mean function, eigenvalues, and eigenvectors coming from an FPCA decomposition of the data, or work using an Anderson-Darling Rank test on FPC scores developed by [60]. Our work focuses on second-

order moment test for functional data. Thus, we focus on tests for covariance operators. Examples include the eigenfunction-based approximation of the distance between two covariance operator proposed by [61]. That work assumed two groups and Gaussian random functions; extensions to non-Gaussian data have been proposed by [62], and for more than two groups proposed by [63].

We now discuss the literature on comparing two covariance matrices in multivariate non-functional settings, as our test relies on comparing the covariance matrix of the FPCA scores. Previous work to compare to covariance matrices the includes likelihood ratio tests [64, 65, 66], as well as tests based on the Frobenius norm proposed by [67, 68]. We based our procedure on work by [69], based on a standardized maximum difference between the entries of two covariance matrices. This test makes no distributional assumptions, has an analytic limiting distribution, and has significantly higher power than other methods in simulation studies.

The rest of the manuscript is structured as follows. Section 2 introduces our novel statistical framework to test the equality of FPC decomposition for two datasets using the covariance matrix of the FPCA scores. Furthermore we extend this framework to paired data, which previous tests have not being developed for and describe the implementation and practical concerns of our methods. Section 3 presents simulation studies designed to explore the statistical performance of our proposed test in various scenarios. Section 4 presents the application of our methodology to our neural-spike data and discusses the scientific implications of our conclusions. We summarize our manuscript and discuss possible future avenues in Section 5.

3.2 Methods

In this section, we develop a test to compare the eigendecompositions of two functional samples. We first introduce a procedure to test two independent samples and proceed to expand to paired functions. The proposed test in this section will be applied to determine whether or not patterns of activation arising in a mouse's motor cortex remain constant over the same experimental conditions.

Define the set of curves $\left[Y_i^{(1)}(t) : i \in \{1, \ldots, I_1\}\right]$ and $\left[Y_i^{(2)}(t) : i \in \{1, \ldots, I_2\}\right]$, where $Y_i^{(1)}(t)$

and $Y_i^{(2)}(t)$ are independent realizations from two random processes $Y^{(1)}(t)$ and $Y^{(2)}(t)$ over $t \in [0,1]$ with corresponding mean function $\mu^{(1)}(t)$ and $\mu^{(2)}(t)$, and covariance operators $\Sigma^{(1)}(s,t)$ and $\Sigma^{(2)}(s,t)$. Furthermore, let $\Sigma^{(z)}(s,t) = \sum_{k=1}^{K} \lambda_k^{(z)} \phi_k^{(z)}(s) \phi_k^{(z)}(t)$ be the truncated spectral decomposition of $\Sigma^{(z)}(s,t)$ for group indicator $z \in \{1,2\}$ coming from Mercer's Theorem, where $\boldsymbol{\Phi}^{(z)}(t) = \left\{\phi_k^{(z)}(t) : k \in \{1,\ldots,K\}\right\}$ are orthonormal eigenfunctions and $\left\{\lambda_k^{(z)} : k \in \{1,\ldots,K\}\right\}$ are the corresponding eigenvalues.

Our goal is to test whether two datasets share the same eigendecomposition, as we aim to compare the shape and scale of eigenfunctions. Equivalently, we can test the covariance operator directly. Thus our goal is to use the two samples of data to test the hypothesis

$$H_0 : \Sigma^{(1)}(s,t) = \Sigma^{(2)}(s,t) \quad \text{versus} \quad H_1 : \Sigma^{(1)}(s,t) \neq \Sigma^{(2)}(s,t). \tag{3.1}$$

We develop methodology to test (3.1) for two independent samples of functions in Section 3.2.1, and extend this procedure to paired functions in Section 3.2.2. We develop our test over a conceptual framework and subsequently describe implementation for real-data settings in Section 3.2.3.

3.2.1 Testing Procedure for Independent Realizations

Under the null hypothesis (3.1) we have, $\boldsymbol{\Phi}(t) = \{\phi_k(t) : k \in \{1,\ldots,K\}\}$, a common set of eigenfunctions and eigenvalues $\boldsymbol{\lambda} = \{\lambda_k : k \in \{1,\ldots,K\}\}$ such that $\Sigma^{(1)}(s,t) = \Sigma^{(2)}(s,t) = \sum_{k=1}^{K} \lambda_k \phi_k(s) \phi_k(t)$. Hence, under the null hypothesis (3.1), a group-specific Karhunen-Loeve expansion of the data is given by

$$Y_i^{(z)}(t) = \mu^{(z)}(t) + \sum_{k=1}^{K} \zeta_{ik}^{(z)} \phi_k(t) \tag{3.2}$$

in which the i^{th} observation in group z is decomposed into a group mean function $\mu^{(z)}(t)$, the common set of eigenfunctions $\boldsymbol{\Phi}(t)$, and the subject-specific scores $\zeta_{ik}^{(z)} = \int_0^1 \{Y_i^{(z)}(t) - \mu^{(z)}(t)\} \phi_k(t) dt$,

assumed to be uncorrelated random variables with mean zero and variance λ_k. We now describe a test for (3.1) using the covariance matrix of the scores coming from the preceding model.

Proposition 1 *Define* $\boldsymbol{\zeta}_i^{(z)} = \left\{\zeta_{ik}^{(z)} : k \in \{1, \ldots, K\}\right\}$ *to be the* $K \times 1$ *random vector of scores derived from (3.2) with mean zero and covariance matrix* $\boldsymbol{\Omega}^{(z)}$ *for subject* $i \in 1, \ldots, I_z$ *in group* $z \in \{1, 2\}$. *Then, the hypothesis test* (3.1) *is equivalent to*

$$H_0 : \boldsymbol{\Omega}^{(1)} = \boldsymbol{\Omega}^{(2)} \quad \textit{versus} \quad H_1 : \boldsymbol{\Omega}^{(1)} \neq \boldsymbol{\Omega}^{(2)}. \tag{3.3}$$

To see that the proposition holds, define the covariance operator for any function $Y_i^{(z)}(t)$ characterized by the model that assumes a shared eigendecomposition (3.2)

$$\text{cov}\left[Y_i^{(z)}(t), Y_i^{(z)}(s)\right] = \sum_{p=1}^{K}\sum_{q=1}^{K} \text{cov}\left(\zeta_{ip}^{(z)}, \zeta_{iq}^{(z)}\right)\phi_p(t)\phi_q(s) \quad z = \{1, 2\} \tag{3.4}$$

because $\Sigma^{(1)}(s,t) = \Sigma^{(2)}(s,t)$ are assumed to be the same under the null, then we have that $\boldsymbol{\Omega}^{(1)} = \boldsymbol{\Omega}^{(2)}$. Hence, the hypothesis (3.1) is equivalent to testing (3.3) using the model that assumes a shared eigendecomposition (3.2).

We compare the covariance matrices of the scores $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$ using work by [69] based on a standardized maximum difference of the two matrices. We test $H_0 : \boldsymbol{\Omega}^{(1)} = \boldsymbol{\Omega}^{(2)}$ by the equivalent hypothesis $H_0 : \max_{1\leq p\leq q\leq K}|\omega_{pq}^{(1)} - \omega_{pq}^{(2)}| = 0$, where $\omega_{pq}^{(z)}$ is the pq^{th} entry of $\boldsymbol{\Omega}^{(z)}$, as the two covariances will only be different if the maximum absolute difference in the same entry of the two covariances is greater than zero. Define $\hat{\omega}_{pq}^{(1)}$ and $\hat{\omega}_{pq}^{(1)}$ to be the pq^{th} entries of sample covariance matrices defined as

$$\hat{\boldsymbol{\Omega}}^{(z)} = \frac{1}{I_z}\sum_{i=1}^{I_z}\left(\boldsymbol{\zeta}_i^{(z)} - \bar{\boldsymbol{\zeta}}^{(z)}\right)\left(\boldsymbol{\zeta}_i^{(z)} - \bar{\boldsymbol{\zeta}}^{(z)}\right)^T \quad \text{where} \quad \bar{\boldsymbol{\zeta}}^{(z)} = \frac{1}{I_z}\sum_{i=1}^{I_z}\boldsymbol{\zeta}_i^{(z)}, \quad z \in \{1, 2\}. \tag{3.5}$$

The test statistic M is the maximum standardized difference in any two entries of the covariance matrices

$$M := \max_{1 \le p \le q \le K} \frac{\left(\hat{\omega}_{pq}^{(1)} - \hat{\omega}_{pq}^{(2)}\right)^2}{\hat{\theta}_{pq}^{(1)}/I_1 + \hat{\theta}_{pq}^{(2)}/I_2} \tag{3.6}$$

where the absolute difference in the same entry is standardized to account for heterogeneity in the estimates of each element of the covariances by

$$\hat{\theta}_{pq}^{(z)} = \frac{1}{I_z} \sum_{i=1}^{I_z} \left[\left(\zeta_{iq}^{(z)} - \bar{\zeta}_p^{(z)} \right) \left(\zeta_{iq}^{(z)} - \bar{\zeta}_q^{(z)} \right) - \hat{\omega}_{pq}^{(z)} \right]^2, \quad z \in \{1, 2\} \tag{3.7}$$

where $\bar{\zeta}_p^{(z)}$ is the p^{th} entry of the vector $\bar{\zeta}^{(z)}$. Under the null distribution that $H_0 : \mathbf{\Omega}^{(1)} = \mathbf{\Omega}^{(2)}$, the value $M - 4\log K + \log\log K$ converges to a Gumbel extreme distribution. The closed form for the rejection threshold at an level α based on K and the appropriate quantile of a Gumbel distribution is $M \le q_\alpha + 4\log K - \log\log K$, where $q_\alpha = -\log(8\pi) - 2\log\log(1-\alpha)^{-1}$ is the $1-\alpha$ quantile of the Gumbel extreme value distribution.

Our tests rely on the asymptotic distribution of the test statistic, and hence we require a sample size that is large enough for the asymptotics to hold. [69] proposed an adjustment to the critical value of the test statistic for small sample sizes and using simulated standard Gaussian datasets to estimate the distribution of the test statistic under the null. We address the case for small sample sizes by introducing a permutation-based empirical null distribution of the test statistic.

3.2.2 Testing Procedure for Dependent or Paired Realizations

We have developed a test for covariance operators from two independent samples of functions that do not directly compares eigenfunctions or eigenvectors and instead compares the covariance of the scores from an FPCA decomposition. We now extend that test to paired data. Consider a samples of paired curves $\left\{ [Y_i^{(1)}(t), Y_i^{(2)}(t)] : i \in \{1, \ldots, I\} \right\}$ that are realizations of the random processes $Y^{(1)}(\cdot)$ and $Y^{(2)}(\cdot)$. Assume that $Y_i^{(1)}(t)$ and $Y_i^{(2)}(t)$ that are repeated observations for the i^{th} subject. Furthermore, consider the sets of subject-specific scores $\left\{ \boldsymbol{\zeta}_i^{(1)} : i = \{1, \ldots, I\} \right\}$ and $\left\{ \boldsymbol{\zeta}_i^{(2)} : i = \{1, \ldots, I\} \right\}$ with covariances $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ coming from the model that assumes

a shared eigendecomposition (3.2). The pairwise correlation that exists among any pairs of functions, will be captured in pairwise correlations between the corresponding pair of subject-specific scores. Thus, we explicitly account for paired functions, by accounting for the pairwise correlation of scores when standardizing our test statistic. Define the updated test statistic for comparing eigendecompositions of paired functions to be

$$M := \max_{1\le p\le q\le K} \frac{\left(\hat{\omega}_{pq}^{(1)} - \hat{\omega}_{pq}^{2}\right)^2}{\left(\hat{\theta}_{pq}^{(1)} + \hat{\theta}_{pq}^{(2)} - 2\hat{\phi}_{pq}\right)/I} \tag{3.8}$$

where $\hat{\omega}_{pq}^{(1)}, \hat{\omega}_{pq}^{(2)}$ are defined as in (3.5), $\hat{\theta}_{pq}^{(1)}, \hat{\theta}_{pq}^{(2)}$ are defined as in (3.7), and

$$\hat{\phi}_{pq} = \frac{1}{I}\sum_{I=1}^{n}\left[\left(\zeta_{ip}^{(1)} - \bar{\zeta}_{p}^{(1)}\right)\left(\zeta_{iq}^{(1)} - \bar{\zeta}_{q}^{(1)}\right)\left(\zeta_{ip}^{(2)} - \bar{\zeta}_{p}^{(2)}\right)\left(\zeta_{iq}^{(2)} - \bar{\zeta}_{q}^{(2)}\right)\right] - \hat{\omega}_{pq}^{(1)}\hat{\omega}_{pq}^{(2)} \tag{3.9}$$

captures the covariance between $\hat{\omega}_{pq}^{(1)}$ and $\hat{\omega}_{pq}^{(2)}$.

The analytic null distribution for the test statistic is not a straightforward computation. We propose a permutation-based approach to estimate the p-value of the test. Define a new version of the original data $\left\{[\tilde{Y}_i^{(1)}(t), \tilde{Y}_i^{(2)}(t)] : i \in \{1, \ldots, I\}\right\}$ such that the label $z = \{1, 2\}$ is randomly permuted within each i^{th} pair, and let $\tilde{M}_p$ to be the test statistic derived from the p^{th} permuted dataset. We construct an empirical null distribution of the test statistic by generating P permuted dataset. Then, the permutation-based test of level α rejects the null hypothesis of equality of eigenfunctions if $M \geq q_{(1-\alpha)}$, where $q_{(1-\alpha)}$ is the $1-\alpha$ quantile of the empirical distribution of the permuted statistics. This permutation scheme provides data under the null distribution that the two groups have the same FPCs while preserving the pairing of the curves.

3.2.3 Practical Implementation

In real-world applications, functions are observed over a vector of discrete grid of timepoints $\boldsymbol{t}_i = \left\{t_{ij} : j \in \{1, \ldots, J_i\}\right\}$ which may vary across subjects. Let $Y_i(t_{ij})$ be the value of $Y_i(\cdot)$ evaluated at t_{ij}, and $Y_i(\boldsymbol{t}_i)$ to be the $J_i \times 1$ vector $Y_i(\cdot)$ evaluated over $\boldsymbol{t}_i$ and assume $Y_i(t_{ij})$ is observed

with some additional random noise with variance σ^2_ϵ. Our goal in this setting is to rely on estimates $\hat{\boldsymbol{\zeta}}_i^{(z)} = \left\{\hat{\zeta}_{ik}^{(z)} : k \in \{1, \ldots, K\}\right\}$ of the true scores in pooled model 3.2 in order to test our hypothesis of equality of eigenfunctions which we derive using a two-step approach. We obtain such estimates from observed data using standard FPCA methods.

In the first step, we estimate the group-specific means of both the groups and remove them so that one can assume the scores have mean zero and can be estimated using off-the-shelf implementations of FPCA. We suggest using a point-wise estimation of the group-specific means for regularly-observed data, and a smoother-based approach using `gam()` in the `mgcv` package [22] for irregularly-observed data. Next, we pool data from both groups after removing group means and estimate shared eigenfunctions, eigenvalues, and subject-specific scores in (3.2) using `fpca.face()` in the `refund` package [13, 41], a fast implementation of standard FPCA. Other FPCA implementation approaches are possible, but we've found this one to be computationally efficient and has worked well in the settings we consider. At last, we perform the two-sample covariance test on the covariance of the estimated scores using the function `HDtest::testCov()` [70].

Our test requires a fixed number of FPCs, just as other leading competing tests do. Selecting a small K or the number of FPCs might lead to a larger type II error, as not enough FPCs will be tested if the difference is in the later eigenfunctions. At the same time, greedy values of K will keep nuisance FPCs that do not provide any information, decreasing power unless the difference is in the later eigenfunctions. We suggest using a percent variance explained (PVE) approach threshold by only keeping the first K components that explain up to some % of the variance. We suggest picking a PVE based on the null hypothesis of interest while interpreting the test in the context of testing a difference among the FPCs that explain % of the variance in the data. We have found that 99% PVE works well in simulations and real data analysis. Furthermore, our simulation studies indicate our test has good power over a reasonable range of PVE thresholds. We suggest additional sensitivity analyses when the test is used in settings not described in our manuscript.

3.3 Simulations

We perform simulation studies to explore our proposed methods' empirical power and size, assessing our test's numerical performance across various scenarios. Across all simulation studies, we compare the performance of our methodology with two leading competing methods, the covariance operator test proposed by [61], and the test proposed in [60]. We first focus on independent datasets and explore the properties over two common real-data analysis scenarios. We explore simulations in which the two groups do not share the same FPCs, scenarios in which both groups have the same set of FPCs, but their eigenvalues differ. We proceed to describe the properties of our test for paired datasets and evaluate them across a range of within-pair correlations.

3.3.1 Simulations for Independent Data

We first generate synthetic datasets for two groups of independent functions. Data is generated from a zero-mean FPCA model in which there are three orthogonal functional principal components defined as

$$Y_i^{(z)}(t) = \sum_{k=1}^{3} \xi_{ik}^{(z)} \sin(2\pi k t) + \epsilon_i(t) \tag{3.10}$$

where we let the scores $\xi_{ik}^{(z)}$ follow a univariate Gaussian distribution with mean zero and a scenario-specific variance. In all simulations, the variance of the first two set of scores is fixed $\text{var}(\xi_{i1}^{(1)}) = \text{var}(\xi_{i1}^{(2)}) = 16$ and $\text{var}(\xi_{i2}^{(1)}) = \text{var}(\xi_{i2}^{(2)}) = 9$. Define $\text{var}(\xi_{i3}^{(1)}) = \gamma+\delta$ and $\text{var}(\xi_{i3}^{(1)}) = \gamma$ given two constants $\delta \geq 0$ and $\gamma \geq 0$. We generate data across various values of baseline variance γ and effect sizes δ. Using this notation, we can explore three distinct scenarios. First, we can estimate the size of the tests when the null hypothesis is true ($\delta = 0, \gamma \geq 0$). Second, we can estimate the power of the test when the two groups do not share the same FPCs ($\delta > 0, \gamma = 0$). Last, we can estimate the power of the test when the two groups share the same FPCs but have different eigenvalues ($\delta > 0, \gamma > 0$). We generate 1000 datasets for each combination of baseline variance $\gamma = (0, 0.1, 0.2, 0.5, 1)$, effect size $\delta = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$, and sample sizes

$I_1 = I_2 = (25, 50, 100, 150, 200)$. We generate data observed over an equally spaced grid $\boldsymbol{t} \in [0, 1]$ of 200 points, and assume that at any j^{th} timepoint the data is observed with noise $\epsilon_i(t_{ij})$, which follows a Gaussian distribution with mean zero and $\sigma_\epsilon^{(2)} = 0.25$.

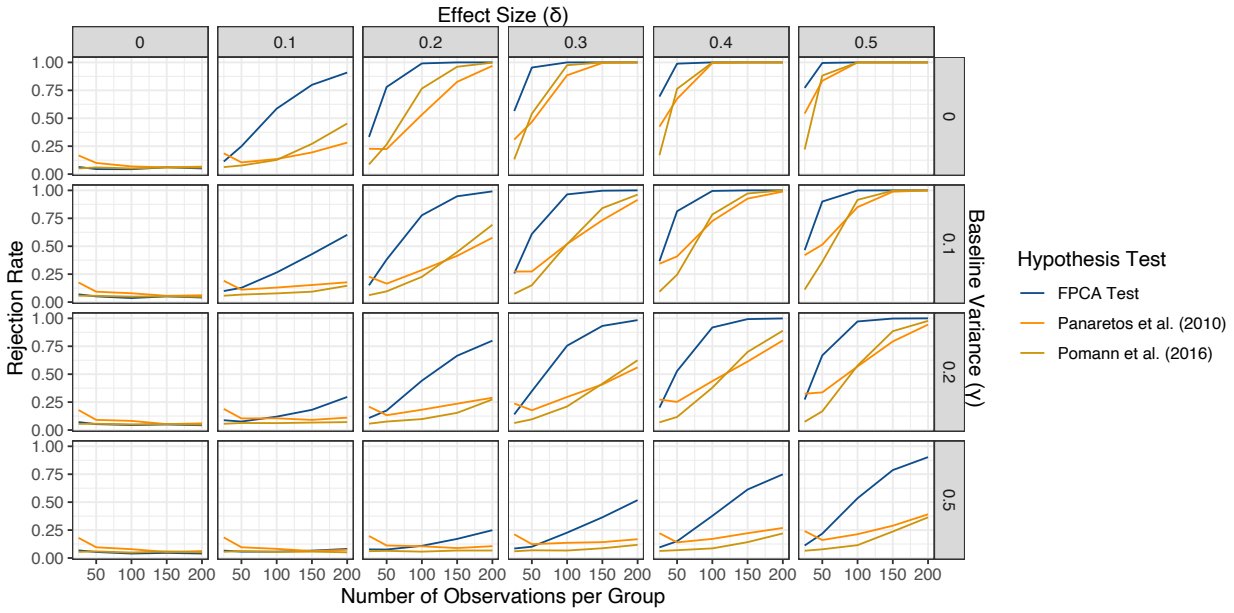


Figure 3.2: Empirical rejection rates across simulation settings. We run 1000 simulations for each simulation scenario, and reject the null hypothesis at $\alpha = 0.05$. Our proposed test is in dark blue. Leading competing methods include the test given in [61] (in orange) and [60] (in yellow). Each column displays a different effect size, and the rows display the baseline variance shared by both groups where $\text{var}(\xi_{i3}^{(1)}) = \gamma + \delta$ and $\text{var}(\xi_{i3}^{(1)}) = \gamma$

For each dataset, we test the null hypothesis at an 5% level. We implement our proposed test following the Section 3.2.3 using a PVE threshold of 99% to select K. We implement the in [61] using the modified variance-stabilized statistic proposed in their manuscript, since simulation studies suggest it provides higher power, and the R function `fdcov::ksample.vstab()` [71]. We follow the guidelines proposed by the authors to select the optimal number of eigenvectors K using the AIC criterion first proposed by [9]. Our procedure estimate the AIC statistic for $K = 2, \ldots, 15$ and select the number of FPCs that yield the minimum value. We implement the [60]'s test using R code provided by the authors and select the number of FPCs in the test by retaining the set of FPCs that explain 99% of the variance. Our supplement provides additional sensitivity analyses that explore the choice of K has on the numerical properties of these tests.

We summarize the numerical evaluation of our proposed method in Figure 3.2. Our simulations show that for small sample sizes, the [61]'s test rejects too frequently; meanwhile, for sample sizes where the test has the right type 1 error, it is less powerful than our method. Our results indicate that our method is more powerful across most scenarios than [60]'s test, and retains the appropriate size when the null is true. Our simulations show that as baseline variance γ increases, it becomes harder for any testing procedures to detect a difference in eigenvalues, suggesting that these methods are more sensitive to scenarios where the two groups don't share the same set of FPCs.

3.3.2 Simulations for Paired Datasets

We generate synthetic datasets with paired data using a variation of the FPCA model (3.10) proposed for independent data samples in the previous section. We introduce non-independence among any two pairs of functions by simulating the scores as correlated multivariate Gaussian variables with a mean of zero rather than independent univariate random variables. In particular, we assume that any k^{th} FPC scores within a pair are correlated and that this pairwise-correlation $\text{corr}(\xi_{ik}^{(1)}, \xi_{ik}^{(2)}) = \rho$ remains constant across the three FPCs. For paired data, allowing $\gamma = 0$ would imply that $cov(\xi_{i3}^{(1)}, \xi_{i3}^{(2)}) = 0$, hence the two samples would be uncorrelated in the FPC in which they differ. Thus, we only focus on analyzing scenarios where the two datasets share the same set of FPCs by fixing $\gamma = 0.5$. We estimate the size of the tests when the null hypothesis is true ($\delta = 0, \gamma = 0.5$) and estimate the power of the test when two samples have a common set of FPCs and different eigenvalues across a variety of effect sizes ($\delta > 0, \gamma = 0.5$).

We generate 100 datasets for each combination of effect size $\delta = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$, baseline variance $\gamma = (0.5)$, and sample size $I = (25, 50, 100, 150, 200)$ and pairwise correlation of the scores $\rho = (0.2, 0.4, 0.6, 0.8)$. We chose these values for ρ as the average correlation between the scores of any given pair in our dataset range between 0.16-0.84. Across all scenarios, we fix $\text{var}(\xi_{i1}^{(1)}) = \text{var}(\xi_{i1}^{(2)}) = 16$ and $\text{var}(\xi_{i2}^{(1)}) = \text{var}(\xi_{i2}^{(2)}) = 9$. We generate data observed over an equally spaced grid $\boldsymbol{t} \in [0, 1]$ of 200 points, and assume that at any j^{th} timepoint the data is observed with noise $\epsilon_i(t_{ij})$, which follows a Gaussian distribution with mean zero and $\sigma_\epsilon^{(2)} = 0.5$.

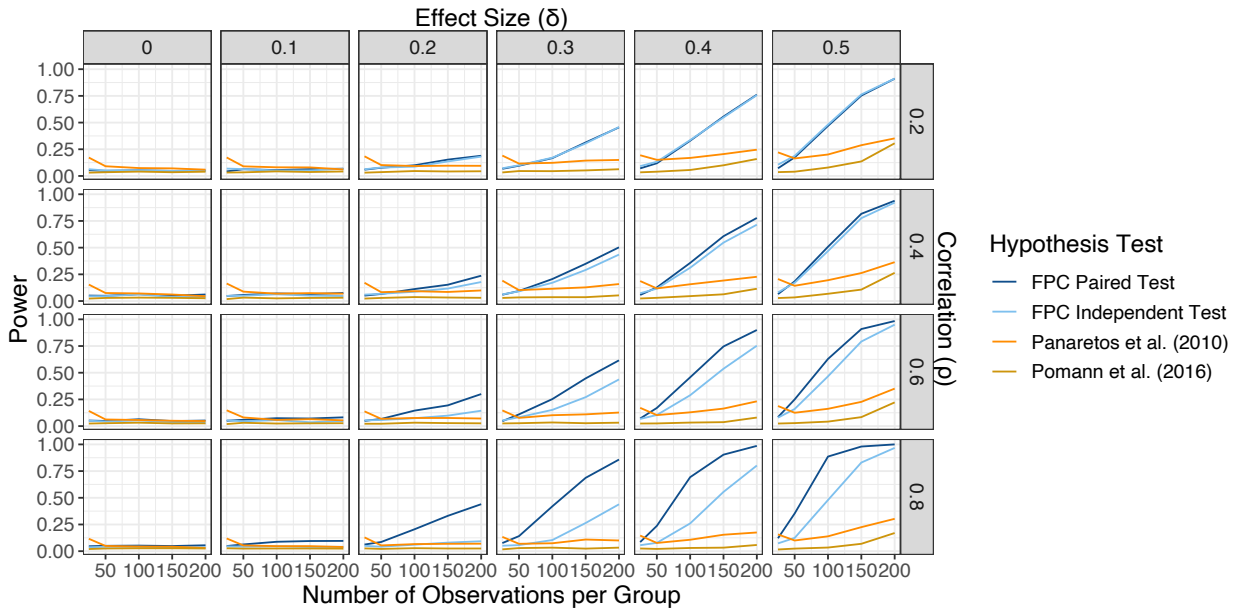


Figure 3.3: Empirical rejection rates across for paired datasets across simulation settings. We run 1000 simulations for each simulation scenario, and reject the null hypothesis at $\alpha = 0.05$. Our proposed paired test is in dark blue. Competing methods include the proposed independent test is in light blue, and the tests given in [61] (in orange) and [60] (in yellow). Each column displays a different effect size, and the rows display the correlation between any two pairs of simulated functions. Across all simulations, $\text{var}(\xi_{i3}^{(1)}) = 0.5 + \delta$ and $\text{var}(\xi_{i3}^{(1)}) = 0.5$

We aim to show how competing methods perform against various degrees of dependence among data pairs and explore how accounting for this dependence leads to a more powerful test. The paired test is compared to the test for independent datasets in section 3.2.1, and the tests given in [61] and [60]. We apply the proposed methodology for comparing FPCs in paired samples as described in section 3.2.2. We implement our paired and independent test as described in Section 3.2.3 using a PVE threshold of 99%. We implement the test given in [61] and [60] in the same way described in the previous section. Our supplement includes sensitivity analyses to show the impact of tuning K across the four competing methods when testing paired functions.

Figure 3.3.2 summarizes the empirical rejection rate across all simulations. Our proposed method for paired data is the most powerful tests when compared to leading competing methods for independent and paired datasets across all the scenarios, including examples in which the set of FPCs in two groups are different and examples in which they are the same, but the eigenvalues

differ. We observe that the improvement in power becomes more pronounced as the pairwise correlation increases. Our test retains the proper rejection rate when the null hypothesis is true.

3.4 Motivating Dataset Analysis

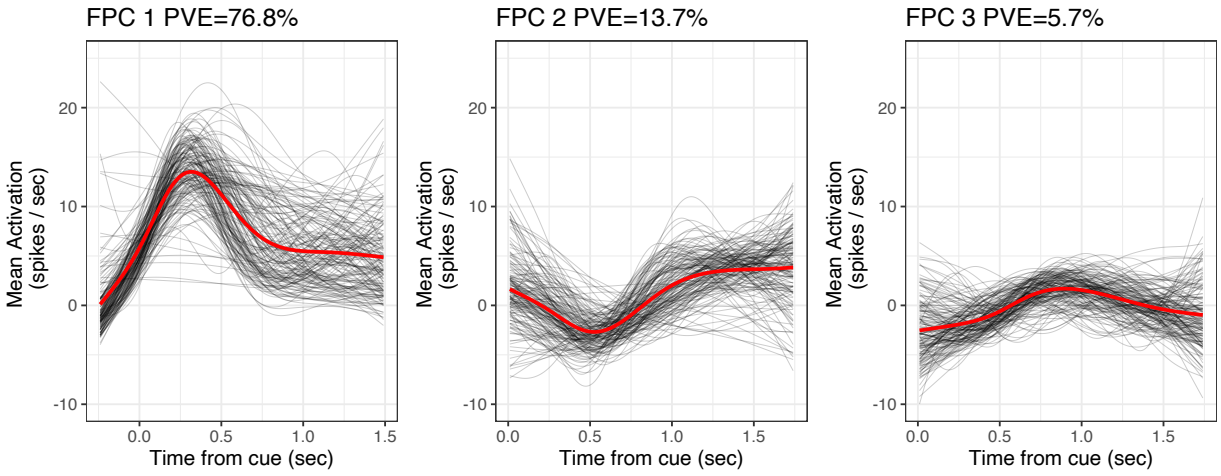


Figure 3.4: Spaghetti plots of FPCA decompositions of trial-level data. Each curve represents an estimate for a trial. The panels show the first three FPCs in descending order of most variance explained. On average, these five FPCs explain 96.2% of the total variability within each trial. The red line is the LOESS average across all trials.

Our goal is to test if activation patterns in the motor cortex remain constant over repeated trials of a trained mouse reaching for a food pellet. To visualize trial-to-trial differences, we begin our real data analysis by fitting FPCA to each trial using `refund::fpca.face`. Figure 3.2 displays spaghetti plots of the first three FPC derived from trial-level decompositions. We show the FPCs in descending order of most variance explained. On average, these 3 FPCs explain 96.2% of the variability in each trial. The red line is the LOESS mean across all trials for each of the individual components. The first FPC explains the intensity of the activation change immediately after the cue; on average it explains 76.8% of the total variance within each trial. The second FPCs explain a trade-off between initial increased activation and lower baseline activation after the cue; on average it explains 13.7% of the total variance within each trial. The third FPC capture local variability unexplained by the first two components; on average, and captures 5.7% of the

remaining variability.

We begin our analysis by applying the test to compare eigendecompositions for paired data developed in Section 3.2.2 to all pairwise comparisons of trials. We summarize the distribution of p-values by calculating the Cramer-Von Mises (CVM) statistic comparing f against the uniform distribution. Let F be the cumulative distribution function of f; then the CVM statistic is defined as $\eta = \int_0^1 [F(x) - x]^2 \, d(x)$. We test the null hypothesis that the activation patterns remain constant by comparing η to an empirical distribution under the null derived using a permutation approach, which we now describe. For each i^{th} Neuron, we permute trial labels, generating data under the null distribution in which there is no effect of trial, yet preserving the repeated measures. Next, we estimate $\tilde{f}_p$, the distribution of p-values for all pairwise comparisons of trials coming from the p^{th} permuted dataset using the test developed in Section 3.2.2. For the p^{th} permutation, calculate $\tilde{\eta}_p = \int_0^1 \left[\tilde{F}_p(x) - x\right]^2 d(x)$ where $\tilde{F}_p$ is cumulative distribution of $\tilde{f}_p$. At last, we reject our null hypothesis at a level α if $\eta \geq q_{(1-\alpha)}$, where $q_{(1-\alpha)}$ is the $1-\alpha$ quantile of the empirical distribution of the permuted statistics $\tilde{\eta} = \left[\tilde{\eta}_p : p \in \{1, \ldots, P\}\right]$. There are a total of 12246 pairwise comparisons comparing 157 trials. We perform 100 permutations to estimate the p-value of each pairwise comparison. We further generate $P = 200$ permuted datasets to generate a global test.

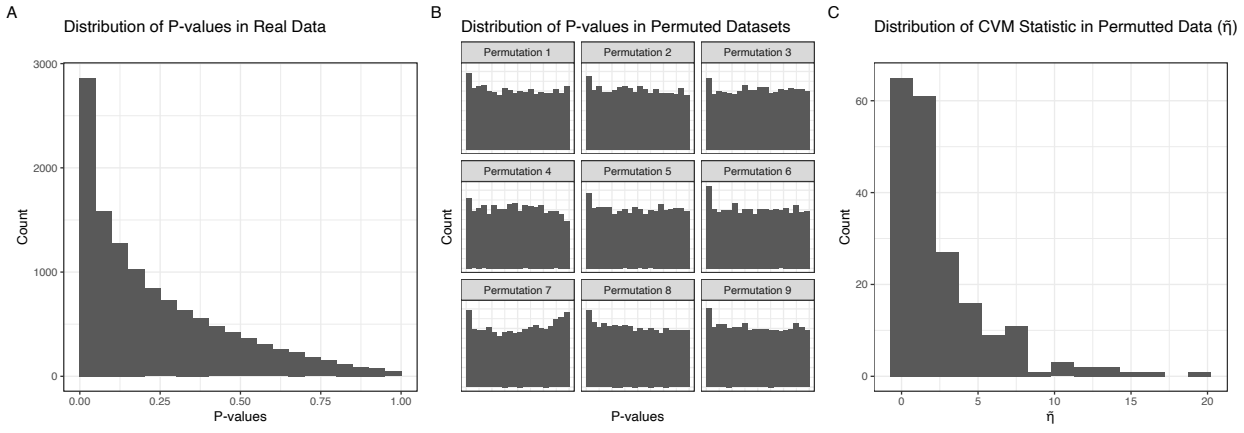


Figure 3.5: Panel A displays the distribution of p-values from all pairwise trial comparisons in our motivating dataset. Panel B shows nine example distributions of p-values from all pairwise trial comparisons in permuted datasets in which the null hypothesis is true. Panel C shows the distribution of $\tilde{\eta}_p$ for $p \in \{1, \ldots, 200\}$.

We summarize the results of our analysis in Figure 3.5. Panel A displays the distribution of the p-values of all pairwise comparisons across all trials for our motivating dataset f. We observe that the distribution is right-skewed, and the test statistic η derived using the observed data is 949.14. Panel B has nine examples of empirical distributions of p-values under the null hypothesis. Each histogram represents a different permuted dataset. Panel C aggregates the results across all permuted datasets and displays the distribution of $\tilde{\eta}$. The range for the empirical distribution is [0.14, 19.52]. We find that the 95% quantile of the distribution is 9.98, and thus we reject the null hypothesis that the activation patterns are constant across trials at a level $\alpha = 0.05$.

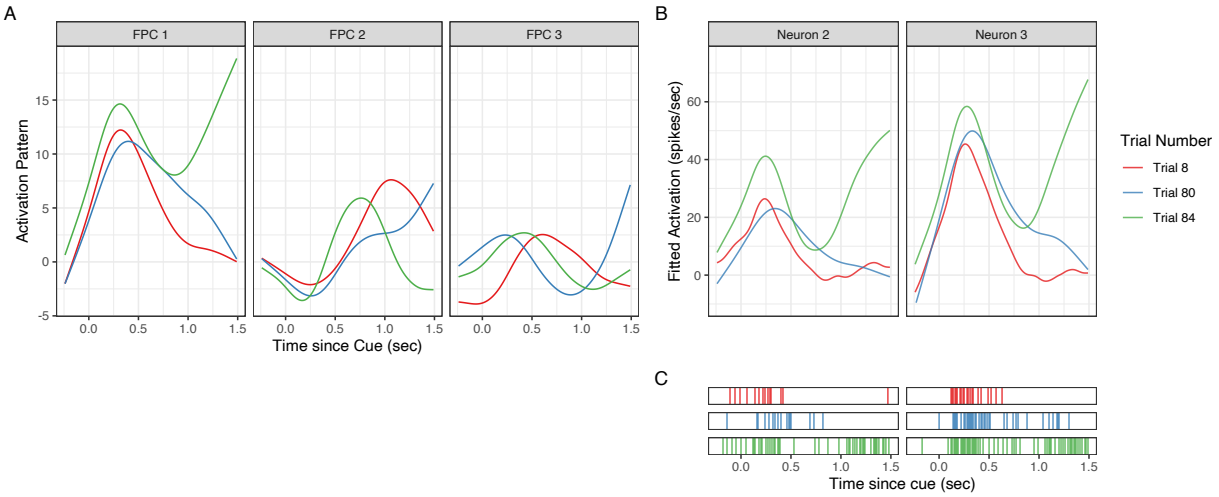


Figure 3.6: Panel A displays the first three activation patterns derived in three trials. Panels B show fitted values in two example neurons previously shown in Figure 3.1. Panel C show barcode plots of raw dichotomized neural spike data. Each colored line represents a timepoint in which that neuron was active.

Figure 3.6 summarizes the activation patterns and neural activation in three example trials. We found that the patterns of activation in Trial 84 are significantly different from those in Trial 8 ($p < 0.01$) and those in Trial 80 ($p < 0.01$). Furthermore, we do not reject the null that the activation patterns differ between Trial 8 and Trial 84 ($p = 0.72$). Panel A summarizes the activation patterns for each of the trials. We observe clear differences in the first activation pattern as there is a consistent late activation across neurons for trial 84. In contrast, the neural activation across neurons decreases for trial 8 and trial 80. Panel B-C shows two example Neurons previously

presented in Figure 3.1. We note that the behavior of the neurons is similar across all trials at the beginning of the reaching movement. However, the neural behavior differs after the 0.75-second mark, as indicated by the derived activation patterns, as both neurons have late activity in Trial 84 while no activity in Trial 8 and Trial 80.

3.5 Discussion

We present a novel approach to test the hypothesis that two sets of functions share a common eigendecomposition based on testing the covariance matrix of the scores coming from an FPCA decomposition. We develop a version of the test for independent functions and another for pairs of functions. We show that our method is more powerful than two leading competing tests and retains the appropriate empirical size when the null hypothesis is true.

We are motivated to understand if activation patterns in neural spike data from a mouse trained to reach for a pellet vary from trial to trial. We analyzed this hypothesis by doing all pairwise comparisons of trials and summarizing these results using the distribution of the p-value. We compare this distribution to the distribution of p-values for all pairwise comparisons under the null distribution using a permutation-based approach. Our results suggest trial-to-trial variability in the activation patterns. These results formally test a hypothesis, a phenomenon described in the neuroscience literature, and provide novel insight into how the motor cortex drives dexterous movement.

A direct extension of our method is extending to the case for K groups. We recognize that such extension could be helpful when testing our global hypothesis that activation patterns remain constant across trials. Furthermore, we could have analyzed our data as counts or dichotomous activation instead of Gaussian. Other potential avenues of exploration include testing the performance of our method for Generalized FPCA to confirm that it remains valid. Our work assumes that both groups are measured over regular discrete grids. We hypothesize that in some scenarios, one of the groups will have higher score variances due to accuracy. Future analysis should explore the performance of our methods when observing the two groups over different structures.

Chapter 4: Last Chapter before conclusion

4.1 Introduction

Between 2001 and 2017, suicide mortality in the United States increased by 31% from 10.7 to 14.0 cases per 100,000 [72]. Previous studies estimate that 8.5%-13% of all suicide attempts are fatal [73, 74, 75] and that around 3% of index attempts lead to death [76]. Roughly half of suicide deaths do not occur during an individual's first time attempt [77, 76]. Thus, preventing non-fatal attempts presents an opportunity for early intervention in a substantial number of people at high risk for suicide [78] and for decreasing the public health burden suicide behaviors.

Despite extensive work over the last 50 years to improve prediction of suicide attempts, a meta-analysis of 365 studies concluded that using known suicide risk factors leads to only slightly better than chance prediction (weighted area under the curve, w/AUC = 0.58). [79] Machine learning methods and big data sources such as electronic health records (EHR) and social media text monitoring have led to substantial improvements in predicting suicide attempts in clinical samples (AUC range 0.71-0.93). [80, 81, 82, 83, 84, 85] However, most of the published literature on non-fatal suicide attempt prediction has focused on high-risk patients who have received mental health treatment. [86, 87]. Over a third of people making non-fatal suicide attempts do not receive mental health treatment [88, 89], and those that engage in mental health treatment only represent one-third of the U.S. fatal suicide attempts [90, 80, 87, 91, 92] These findings underscore the importance of extending suicide prediction models beyond high-risk populations to the general adult population.

In the present study, we aim to identify important risk factors for future suicide attempt in the general population by taking advantage of the richness of the NESARC dataset using an explanatory machine learning model. We extend prior research in 3 important directions: 1) We use a

large, nationally representative longitudinal sample, to identify risk factors for suicide attempt in the general population; 2) We use an extensive assessment instrument that includes detailed evaluation of substance use and psychiatric disorders and symptoms that are not routinely available in EHR or administrative data; and 3) We incorporate class imbalance as a feature in our model to address the limitations of more generic algorithms, as few studies have previously done this [87]. Overall, we expected to confirm previously identified risk factors found in clinical samples and, more importantly, to identify new ones to expand our understanding of the etiology of suicide attempts.

4.2 Methods

4.2.1 Sample

Data were drawn from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), a face-to-face survey conducted with a nationally representative sample of the U.S. adult population by the National Institute on Alcoholism and Alcohol Abuse (NIAAA) [93]. The target population included the civilian noninstitutionalized population, aged 18 years and older, in the United States. Wave 1 NESARC survey data (2001-2002) and self-reported non-fatal suicide attempts at follow up 3 years later (Wave 2, 2004-2005) [94]. . The cumulative response rate at Wave 2 was 70.2%, resulting in 34,653 Wave 2 interviews. Survey design and non-response weights are available to allow estimates to be representative of the U.S. civilian population based on the 2000 census. [95]. The research protocol, including written informed consent procedures, received full human subjects review and approval from the U.S.Census Bureau and the Office of Management and Budget.

4.2.2 Predictors from Wave 1

At Wave 1, participants were assessed using the Alcohol Use Disorder and Associated Disabilities Interview Schedule DSM-IV version (AUDADIS-IV) [93, 96], , a lay-administered structured interview to assess alcohol, drug, and mental health disorders according to DSM-IV criteria. Past

12 months axis I disorders evaluated included substance use disorders (alcohol use, drug use, and nicotine dependence), mood disorders (major depressive, dysthymic, and bipolar disorder), anxiety disorders (panic disorder, social anxiety disorder, specific phobia, and generalized anxiety disorder) and pathological gambling. Axis II disorders included avoidant, dependent, obsessive-compulsive, histrionic, paranoid, schizoid, and antisocial personality disorders assessed on a lifetime basis. Demographic and background information was collected. Response patterns for each of the 14 sections of the survey are summarized in the Supplemental eMethods B.1.The test-retest reliability of AUDADIS-IV and its validity for measuring DSM-IV mental disorders is good to excellent for substance use (κ=0.51-0.74) and fair to good for other disorders (κ=0.40-0.67). [97, 98, 99, 100]

The Wave 1 survey contained 2,805 separate questions. To reduce interview burden, participants skipped entire sections based on their responses to gate questions. Additionally, there were 180 derived variables that include DSM-IV past year, prior to past year, and lifetime diagnoses of mental disorders including personality disorders. For each Wave 1 participant, there were between 643 and 2,985 available features.

4.2.3 Outcome at Wave 2: Non-fatal suicidal attempt

At Wave 2, a similar face-to-face structured interview follow up was conducted. The primary outcome was having attempted suicide in the previous 3 years since the time of the Wave 2 interview. This variable was derived by combining responses to the Wave 2 questions: “In your entire life, did you ever attempt suicide?” and if affirmative, answers to “How old were you the first time?” and “How old were you the most recent time?”. [101] If the most recent suicide attempt occurred within the last 3 years the participant was considered a case, otherwise not. At Wave 2, 222 participants endorsed having attempted suicide since the Wave 1 interview.

4.2.4 Data Analysis

Model Building

We performed an initial data analysis as recommended [102], and addressed the survey structural missingness by using the missing-indicator method [103, 104] as described in the supplement. We used balanced random forest (BRF) to build a model to identify factors associated with suicide attempts by taking the processed 2,978 Wave 1 features to classify dichotomous suicide attempt at Wave 2. BRF has better performance than regular random forest for classification models with class-imbalanced data [105, 106]. As detailed in the Supplemental eMethods B.1, we tuned the BRF parameters by using 10-fold cross-validation, and further validated our classification model by using nested cross-validation [107].

We summarized the final model's performance by aggregating the out-of-fold classifications of our optimal model and used this aggregated probability ("threshold") to calculate an out-of-fold AUC. We weighted our results based on design and non-response weights to allow our estimates to be representative of the U.S. civilian population based on the 2000 census. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), alarms per 100 evaluations, and number needed to evaluate to find one new suicide attempt case were examined against the threshold value.

Identifying Top Risk Predictors

To quantify model variable importance, we calculated the decrease in classification ability after any individual feature was permuted (i.e., no longer used in the suicide attempt model) across the dataset [108, 109]. The importance measure was scaled between 0 and 100 by subtracting the smallest importance from all observations and dividing by the largest importance. To facilitate interpretation of suicide attempt risk, we defined 4 interpretable risk severity groups that could be used as a reference of suicide attempt subgroups in the U.S. adult population based on the BRF model, as defined in the Supplemental eMethods B.1. We calculated summary statistics of suicide

attempt broken down by risk groups. Finally, we quantified the risk associated with each of the top-performing risk factors by generating response plots of the distribution of probabilities for all observations in the four empirical-derived risk groups.

Model Validation

We further validated our model by 1) calculating classification performance stratified by time-to-suicide attempt from the first interview; 2) stratifying classification performance across sex, age, self-reported race (white vs. non-white), and income groups to test the robustness against demographic characteristics; and, 3) examining erosion in model accuracy with fewer features by running additional BRFs using only the top 5 or 10 risk factors selected from the random forest importance measure.

4.3 Results

4.3.1 Performance of the Suicide Prediction Model

We found that 0.64% (N=222) participants attempted suicide out of 34,653. The out-of-sample AUC for the best model, including all Wave 1 features, was 0.857 (range 0.803-0.909). The optimal cross-validated number of variables to sample at each fold was 1,700, representing 57.1% of all features. The out-of-sample generalizability, defined as the correlation between our final model and our nested cross-validated model, was 0.997. Figure 4.1 presents the distribution of model-calculated risk scores across the whole sample, and stratified by whether participants reported a suicide attempt at Wave 2. Suicide attempt risk strata (low, medium, high, very high) are summarized in Table 4.1. Base on our model, 73.1% of the U.S. population is at low risk, 17.5% is at medium risk, 7.6% is at high risk, and 1.8% is at very high risk of suicide attempt. Based on these categories, 62.2% of individuals (138 of 222) who attempted suicide between Waves 1 and 2 were in the high or very high risk groups based on their Wave 1 survey responses. Thirty-two attempters (14.4% of cases) were classified as being low risk.

Figure 4.2 displays sensitivity, specificity, PPV, NPV, alarms per 100 evaluations, and number

needed to evaluate to find one new suicide attempt case across various classification thresholds. For each plot, we labeled the 3 thresholds used to classify the 4 risk groups. If using the very-high-risk group as a classification threshold, there would be 2 alarms, an NNE of 10, a PPV of 10.4%, and an NPV of 99.6%. If using the high-risk group, or the top decile of risk, as a threshold for classification, there would be 10 alarms, an NNE of 26, a PPV of 3.9%, and an NPV of 99.7%. Using the threshold that optimizes Youden's statistic, there would be 27 alarms, an NNE of 51, a PPV of 2.0%, and an NPV of 99.9%.

4.3.2 Variable Importance and Risk Factor Effects

Table 4.2 shows the 20 most important variables from the BRF model. The 3 most important risk factors were "felt like wanted to die," "thought about committing suicide," and previous suicide attempt. Several of the most important variables were associated with past month low energy and mood periods, such as feeling downhearted, feeling less accomplished, or paying less attention to work or other activities. Other features identified as important were age, family income and financial crisis, marital status, education level, paternal alcohol problem, and parental separation. Supplemental eFigure B.1 shows the distribution of model-calculated scores as a function of the top 10 most important variables identified by the BRF algorithm allowing interpretation for how each variable was associated with suicide attempt.

4.3.3 Model Robustness Results

We conducted a series of sensitivity and complementary analyses. Supplemental eTable B.2, shows that the classification ability of the model decreased over time-to-suicide attempt from the first interview. The percentage of participants classified as very-high-risk was 50.46% for those who attempted suicide within the first year, 33.11% for those who attempted suicide between the first and second year, 30.30% for attempters between the second and third year, and 16.68% for those who attempted between the third and follow-up. Second, we examined the classification ability of our model across demographic characteristics. As shown in Supplemental eTable B.4,

the AUC was 0.808 (CI=0.765-0.851) for participants aged 18-36 years, 0.867 (CI=0.827-0.906) for those aged 37-53, and 0.872 (CI = 0.8-0.945) for those 54 years old or older. We found that the AUC was 0.850 (CI=0.813-0.886) for females (57.97%) and 0.872 (CI=0.84-0.904) for males. The AUC was 0.877 (CI=0.845-0.909) for white participants and 0.831 (CI=0.788-0.873) for non-white participants. Finally, the AUC was 0.845 (CI=0.813-0.877) for participants with an income lower than $20,000, 0.848 (CI=0.786-0.91) for those with an income between $20,000-34,999, 0.794 (CI=0.676-0.911) for those with incomes between $35,000-69,999, and 0.944 (CI=0.893-0.994) for those with incomes higher than $70,000.

At last, we measured the decrease in model accuracy when fewer features were included by building new separate BRF models with the top 5 and 10 most important variables previously found using the entire feature set. The out-of-sample AUCs for these models were 0.818 (SE = 0.017) and 0.845 (SE = 0.016), respectively.

4.4 Discussion

We built a model to classify non-fatal suicide attempts using a large, nationally representative sample of U.S. adults. It confirmed several well-known risk factors of suicide attempt and identified several new ones. When tested outside the training set, our model performed at levels similar to models restricted to data from high-risk mental health patients [79, 84] for the full sample and when stratified by demographic characteristics, indicating its robustness. Its classification power decreased with time elapsed from the baseline interview, providing an indirect measure of its validity. These results are encouraging given the recent emphasis on models in the general adult population using rich dataset, and their usefulness to develop precision treatment rules for a suicide attempt [110, 87, 111]

We found significant conceptual overlap of our most important risk factors with items commonly used in suicide risk scales. In accord with previous findings,[112, 113, 114, 115, 116]the strongest risk factors for future suicide attempts were related to previous suicidal behaviors. For example, “felt like wanted to die” is covered in the Patient Health Questionnaire (PHQ-9) [117],

the Columbia Suicide Severity Rating Scale (C-SSRS) [118], and Beck's Scale of Suicide Ideation (BSSI). [119] Previous suicide ideation is covered in the BSSI, the SAD PERSONS Scale [120], and the Suicide Assessment Scale (SUAS) [121], while previous suicide attempt is covered in the SAD PERSONS scale and the C-SSRS. Feeling downhearted and depressed is covered in the PHQ-9 [117] (item #2) as well as the Beck Depression Inventory (item #1). [122, 119]

Our results extend prior work by revealing the predictive value of variables related to functional impairment resulting from mental disorders, which are not generally covered in screening tools for suicide risk assessment. The questions identified from the 12-Item Short Survey (SF-12) [123, 124, 125] impairment constructs tapped accomplishing less than you like and not performing work or other activities as carefully as usual. These findings may offer new avenues to improve suicidal behavior prediction through functional assessment. We note that sex was not one of the most important variables, suggesting that sex differences in other risk factors are likely to mediate the difference in suicide attempt prevalence across sexes [126, 78, 127].

Other important novel risk factors identified were related to socio-economic disadvantage. Specifically, lower educational level, and experiencing a financial crisis in the last year were among the 10 most important variables. Seeking to alleviate the economic and emotional effects of financial crises might be an important aspect of suicide risk prevention, particularly in the context of deaths of despair formulations of suicide risk [128, 129], This is of particular contemporary relevance, given increased unemployment and economic stress in the United States related to the COVID-19 pandemic [130]. Our study identifies an individual-level association between economic strain with suicide attempt risk that extend beyond findings of previous studies showing: 1) population-level relationship between economic recessions and increased suicide rates [131, 132, 133]; 2) a link between financial debt and suicide ideation [134]; and 3) case-control research linking unemployment and personal debt to suicide risk [135]. Although this association has previously been reported in the NESARC [136], our data-driven results highlight this risk factor as one of the most important for suicide attempt in the general population.

We incorporated technical advances in the modeling by using balanced random forest to ad-

dress the extreme class imbalance and by using the missing indicator approach to address gate questions and skip patterns common to survey data. We ensured population-level generalizability by incorporating the complex survey design and sampling weights. The algorithms in this study may be useful for the analysis of other large survey datasets. Our methods may have wide applications, given the NIH's recent decision to link research samples to the National Death Index [137], and the greater availability of longitudinal mortality outcomes for cross-sectional surveys.

This study has some limitations. First, we only have data from participants who were aged 18 years old and over, and some of the risk factors identified, such as financial crisis, might only be relevant to adult populations. Furthermore, suicide risk is highest for people aged 15 to 25 years [138]. Second, we do not have information about suicide attempts among subjects lost to follow up, i.e. Wave 2 non-responders including participants who died of suicide, which would have enhanced our ability to detect differences between fatal and non-fatal suicide attempts. Nevertheless, we found lower rates of prior suicidal behaviors and ideation at Wave 1 among Wave 2 non-responders suggesting that selection bias related to suicide attempts is likely small (Supplemental eTable B.3). Third, there is potential for misclassification of suicide attempt. There is uncertain reliability of self-reported suicide attempt over such a long recall period, and the willingness to disclose previous attempts in a face-to-face interview [139, 140].However, our findings confirming previous risk factors add face validity to our results. Fourth, our study examined occurrence of suicide attempts within three years of assessment. Exploration of shorter and more clinically relevant time horizons should also be evaluated. Furthermore, the association between risk factors and future suicide attempts may vary over time. Fifth, the data were collected in 2000-2005, and there may have been recent secular changes in risk factors for suicide attempts. Sixth, the survey was not collected to study suicide, and some important covariates, such as stress and adjustment disorders, were not included. Furthermore, Wave 1 suicide symptom was only asked of participants who endorsed depressed mood or anhedonia. Given the important role this item assumed as a risk factor of future suicide attempt, an item that was asked of everyone might have increased the accuracy of the models.

4.5 Conclusion

Our study demonstrates the ability of machine learning methods to generate powerful and parsimonious suicide attempt models in general adult population samples that build on and complement knowledge derived from clinical and high-risk samples. We confirmed several well-known risk factors of suicide attempts, such as previous suicidal behaviors and depression, while identifying new important risks. Specifically, functional impairment and socio-economic disadvantage emerged as novel important factors of suicide attempt in the general population with lower education educational level and recent financial crisis as an individual-level risk of future suicide attempts. We hope that these results deepen our understanding of the etiology of suicide attempts in adults and improve suicidal behavior prediction by identifying new risk variables to guide clinical assessment and development of suicide risk scales.

Table 4.1: Summary of suicide attempt risk during the first three years after wave one NESARC[a] interview weighted to be representative of the U.S. population[c] and broken down by predicted riskc group based on wave one responses (P = 2985)

Risk Group[b]	Total	Proportion of total sample	N Attempted	N Did not attempt	General Population PPV of attempt in next 3 years [b]	General Population NPV of no attempt in next 3 years [b]	In sample PPV of attempt in next 3 years	In sample NPV of no attempt in next 3 years	Mean Predicted Score[c]
low	24862	71.7%	32	24830	0.0013	0.9987	0.0013	0.9987	16.2
medium	6116	17.7%	52	6064	0.0084	0.9916	0.0085	0.9915	41.8
high	2945	8.5%	69	2876	0.0244	0.9756	0.0234	0.9766	61.6
very high	730	2.1%	69	661	0.1003	0.8997	0.0945	0.9055	81.3
Total	34653	100%	222	34431	0.006	0.9937	0.006	0.9936	25.5

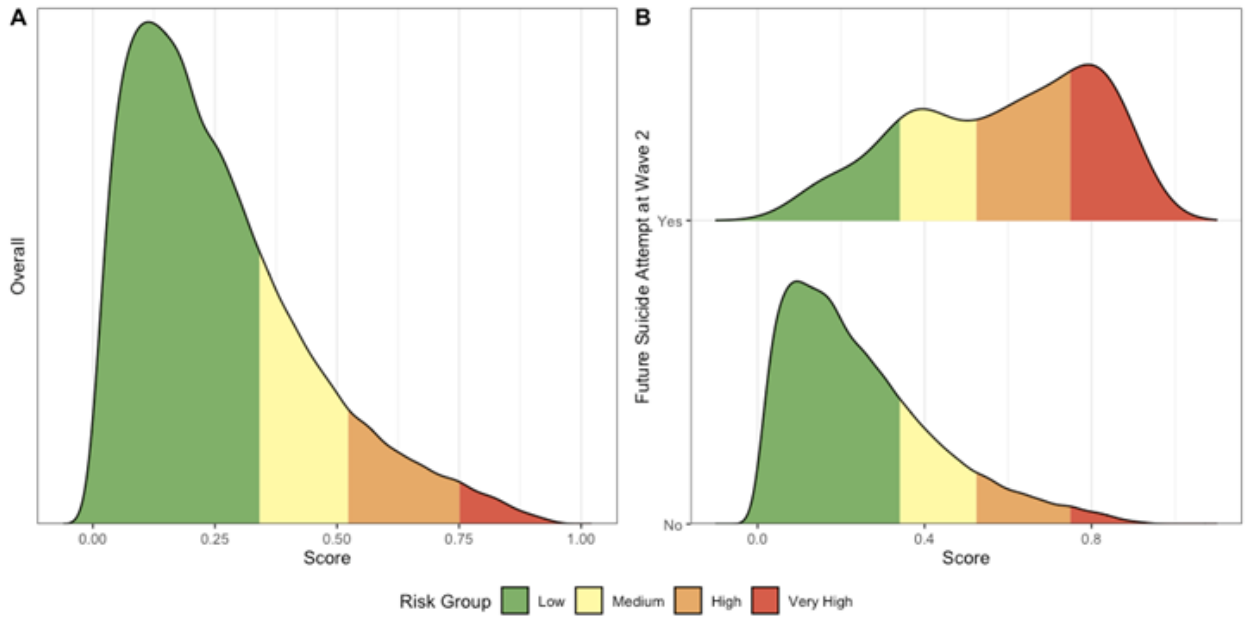
Footnote:

[a] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005.

[b] Data is weighted to be representative of the U.S. adult population for region, age, sex, race, and ethnicity, based on the 2000 Census

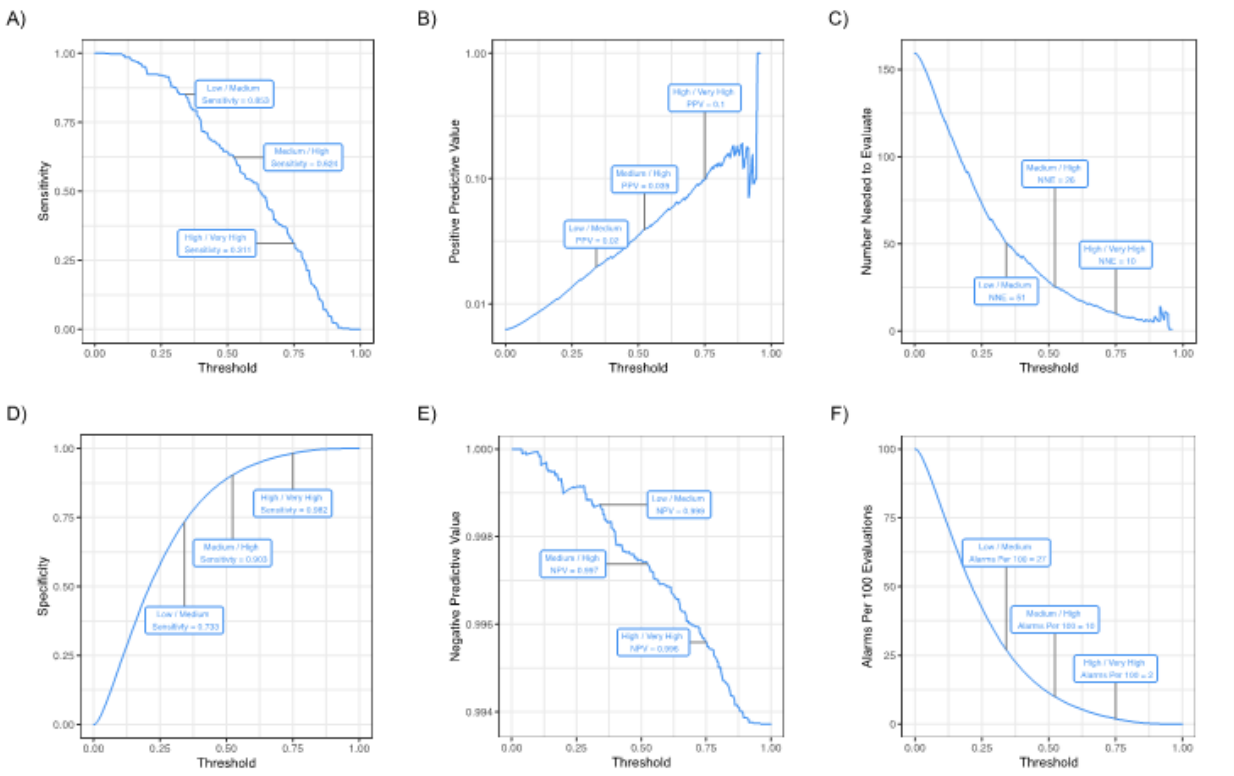
[c] Predictions are based on the out-of-sample prediction from our machine learning balanced random forest model

Figure 4.1: Distribution of the predicted risk scores [a] based on NESARC [b] wave one responses (P = 2985) weighted to be representative of the U.S. population [c] colored by predicted risk group [d] A) Distribution of Scores across the entire sample B) Distribution of scores broken down by cases (N = 222) and controls (N = 34431)



Footnote:
[a] Predictions are based on the out-of-sample prediction from our machine learning balanced random forest model [b] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005.
[c] Data is weighted to be representative of the U.S. adult population for region, age, sex, race, and ethnicity, based on the 2000 Census
[d] The predicted risk groups are defined on interpretable thresholds for suicide risk prediction.

Figure 4.2: Summary measures of the predictive ability of the model based on NESARC [a] wave one responses (P = 2985) calculated across all possible classification thresholds. The highlighted cutoffs are the those used to define our risk prediction groups. [b] Results were weighted to be representative of the U.S. population [c]



Footnote:

[a] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005.

[b] The predicted risk groups are defined on interpretable thresholds for suicide risk prediction.

[c] Data is weighted to be representative of the U.S. adult population for region, age, sex, race, and ethnicity, based on the 2000 Census

Table 4.2: Top 20 most important variables based on our suicide prediction model using NESARC [a] wave one responses (P = 2985).

	Description	Importance Score[b]
1	Felt like wanted to die	100.000
2	Thought about committing suicide	48.425
3	Attempted suicide	21.932
4	During past 4 weeks, how often felt downhearted and depressed	14.033
5	Age	13.731
6	During past 4 weeks, how often did work or other activities less carefully than usual as result of emotional problems	13.051
7	Experienced major financial crisis, bankruptcy or been unable to pay bills on time in last 12 months	11.478
8	During past 4 weeks, how often accomplished less than would like as result of emotional problems	11.213
9	Grade level during 2000-2001 school year	10.319
10	Highest grade or year of school completed	7.938
11	During past 4 weeks, how often physical health or emotional problems interfered with social activities	7.746
12	Blood/natural father ever an alcoholic or problem drinker	7.377
13	Occupation: current or most recent job	6.059
14	Current marital status	4.727
15	Family income in last year	4.472
16	Age when biological/adoptive parents stopped living together	4.471
17	Thought a lot about own death	4.135
18	Present situation includes in school part time	4.128
19	Personal income in last year	4.122
20	Parent lived with after biological or adoptive parents stopped living together	4.037

Footnote:

[a] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005.

[b] The importance score was calculated by permuting the labels and estimating the decrease in prediction

References

[1] J. Ramsay and B. W. Silverman, *Functional Data Analysis* (Springer Series in Statistics), 2nd ed. New York: Springer-Verlag, 2005, ISBN: 978-0-387-40080-8.

[2] B. A. Sauerbrei *et al.*, "Cortical pattern generation during dexterous movement is input-driven," *Nature*, vol. 577, no. 7790, pp. 386–391, 2020.

[3] P. Besse and J. O. Ramsay, "Principal components analysis of sampled functions," *Psychometrika*, vol. 51, no. 2, pp. 285–311, 1986.

[4] J. O. Ramsay and C. Dalzell, "Some tools for functional data analysis," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 53, no. 3, pp. 539–561, 1991.

[5] J. A. Rice and B. W. Silverman, "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 53, no. 1, pp. 233–243, 1991.

[6] S Pezzulli and B Silverman, "Some properties of smoothed principal components analysis," *Computational Statistics*, vol. 8, pp. 1–16, 1993.

[7] B. W. Silverman *et al.*, "Smoothed functional principal components analysis by choice of norm," *The Annals of Statistics*, vol. 24, no. 1, pp. 1–24, 1996.

[8] G. Boente and R. Fraiman, "Kernel-based functional principal components," *Statistics & probability letters*, vol. 48, no. 4, pp. 335–345, 2000.

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

[10] P. Hall, H.-G. Müller, and J.-L. Wang, "Properties of principal component methods for functional and longitudinal data analysis," *The annals of statistics*, pp. 1493–1517, 2006.

[11] C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi, "Multilevel functional principal component analysis," *The annals of applied statistics*, vol. 3, no. 1, p. 458, 2009.

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

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

[14] L. Xiao, C. Li, W. Checkley, and C. Crainiceanu, "Fast covariance estimation for sparse functional data," *Statistics and computing*, vol. 28, no. 3, pp. 511–522, 2018.

[15] M. E. Tipping and C. M. Bishop, "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999, _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00196.

[16] G. M. James, T. J. Hastie, and C. A. Sugar, "Principal component models for sparse functional data," *Biometrika*, vol. 87, no. 3, pp. 587–602, 2000.

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

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

[19] A. E. Hoerl and R. W. Kennard, "Ridge regression: Applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.

[20] P. J. Brown and J. V. Zidek, "Adaptive multivariate ridge regression," *The Annals of Statistics*, vol. 8, no. 1, pp. 64–74, 1980.

[21] K.-C. Li, "Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing," *The Annals of Statistics*, pp. 1101–1112, 1986.

[22] S. N. Wood, "Mgcv: Gams and generalized ridge regression for r," *R news*, vol. 1, no. 2, pp. 20–25, 2001.

[23] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric regression*. Cambridge university press, 2003, Issue: 12, ISBN: 0-521-78516-2.

[24] M. P. Wand, "A comparison of regression spline smoothing procedures," *Computational Statistics*, vol. 15, no. 4, pp. 443–462, 2000, ISBN: 0943-4062.

[25] G. Wahba, *Spline models for observational data*. SIAM, 1990.

[26] C. Gu, "Adaptive spline smoothing in non-gaussian regression models," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 801–807, 1990.

[27] A. Pintore, P. Speckman, and C. C. Holmes, "Spatially adaptive smoothing splines," *Biometrika*, vol. 93, no. 1, pp. 113–125, 2006.

[28] C. Gu and C. Gu, *Smoothing spline ANOVA models*. Springer, 2013, vol. 297.

[29] D. Ruppert and R. J. Carroll, "Spatially-adaptive Penalties for Spline Fitting," *Australian & New Zealand Journal of Statistics*, vol. 42, no. 2, p. 205, Jun. 2000, Publisher: Wiley-Blackwell.

[30] V. Baladandayuthapani, B. K. Mallick, and R. J. Carroll, "Spatially Adaptive Bayesian Penalized Regression Splines (P-splines)," *Journal of Computational and Graphical Statistics*, vol. 14, no. 2, pp. 378–394, Jun. 2005, Publisher: Taylor & Francis.

[31] T. Krivobokova, C. M. Crainiceanu, and G. Kauermann, "Fast Adaptive Penalized Splines," *Journal of Computational and Graphical Statistics*, vol. 17, no. 1, pp. 1–20, Mar. 2008, Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/106186008X287328.

[32] Z. Liu and W. Guo, "Data Driven Adaptive Spline Smoothing," *Statistica Sinica*, vol. 20, no. 3, pp. 1143–1163, 2010, Publisher: Institute of Statistical Science, Academia Sinica.

[33] A. Petersen, D. Witten, and N. Simon, "Fused lasso additive model," *Journal of Computational and Graphical Statistics*, vol. 25, no. 4, pp. 1005–1025, 2016.

[34] O. H. Madrid Padilla, J. Sharpnack, Y. Chen, and D. M. Witten, "Adaptive nonparametric regression with the k-nearest neighbour fused lasso," *Biometrika*, vol. 107, no. 2, pp. 293–310, 2020.

[35] Y. Grandvalet, "Least absolute shrinkage is equivalent to quadratic penalization," in *International Conference on Artificial Neural Networks*, Springer, 1998, pp. 201–206.

[36] S Canu and Y Grandvalet, "Outcomes of the equivalence of adaptive ridge with least absolute shrinkage," *Advances in neural information processing systems*, p. 445, 1999.

[37] F. Frommlet and G. Nuel, "An adaptive ridge procedure for l 0 regularization," *PloS one*, vol. 11, no. 2, e0148620, 2016.

[38] L. Dai, K. Chen, and G. Li, "The broken adaptive ridge procedure and its applications," *Statistica Sinica*, vol. 30, no. 2, pp. 1069–1094, 2020.

[39] M. P. Wand and J. T. Ormerod, "On Semiparametric Regression with O'sullivan Penalized Splines," *Australian & New Zealand Journal of Statistics*, vol. 50, no. 2, pp. 179–198, 2008, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-842X.2008.00507.x.

[40] T. Speed, "[that blup is a good thing: The estimation of random effects]: Comment," *Statistical science*, vol. 6, no. 1, pp. 42–44, 1991.

[41] J. Goldsmith *et al.*, "Refund: Regression with functional data," *R package version 0.1-16*, vol. 572, 2016.

[42] B. A. Sauerbrei *et al.*, "Cortical pattern generation during dexterous movement is input-driven," *Nature*, vol. 577, no. 7790, pp. 386–391, 2020.

[43] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: State-of-the-art and future challenges," *Nature neuroscience*, vol. 7, no. 5, pp. 456–461, 2004.

[44] K. L. Briggman, H. D. Abarbanel, and W. Kristan Jr, "Optical imaging of neuronal populations during decision-making," *Science*, vol. 307, no. 5711, pp. 896–901, 2005.

[45] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[46] B. M. Yu, J. P. Cunningham, G. Santhanam, S. Ryu, K. V. Shenoy, and M. Sahani, "Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity," *Advances in neural information processing systems*, vol. 21, pp. 1881–1888, 2008.

[47] B. M. Broome, V. Jayaraman, and G. Laurent, "Encoding and decoding of overlapping odor sequences," *Neuron*, vol. 51, no. 4, pp. 467–482, 2006.

[48] G. Santhanam *et al.*, "Factor-analysis methods for higher-performance neural prostheses," *Journal of neurophysiology*, vol. 102, no. 2, pp. 1315–1330, 2009.

[49] M. M. Churchland *et al.*, "Stimulus onset quenches neural variability: A widespread cortical phenomenon," *Nature neuroscience*, vol. 13, no. 3, pp. 369–378, 2010.

[50] K. V. Shenoy, M. Sahani, M. M. Churchland, *et al.*, "Cortical control of arm movements: A dynamical systems perspective," *Annu Rev Neurosci*, vol. 36, no. 1, pp. 337–359, 2013.

[51] C. Hartmann, A. Lazar, B. Nessler, and J. Triesch, "Where's the noise? key features of spontaneous activity and neural variability arise through learning in a deterministic network," *PLoS computational biology*, vol. 11, no. 12, e1004640, 2015.

[52] R. D. Flint, M. R. Scheid, Z. A. Wright, S. A. Solla, and M. W. Slutzky, "Long-term stability of motor cortical activity: Implications for brain machine interfaces and optimal feedback control," *Journal of neuroscience*, vol. 36, no. 12, pp. 3623–3632, 2016.

[53] J. A. Gallego, M. G. Perich, R. H. Chowdhury, S. A. Solla, and L. E. Miller, "Long-term stability of cortical population dynamics underlying consistent behavior," *Nature neuroscience*, vol. 23, no. 2, pp. 260–270, 2020.

[54] C. Zhang, H. Peng, and J.-T. Zhang, "Two Samples Tests for Functional Data," *Communications in Statistics - Theory and Methods*, vol. 39, no. 4, pp. 559–578, Feb. 2010, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610920902755839.

[55] L. Horváth and P. Kokoszka, *Inference for functional data with applications*. Springer Science & Business Media, 2012, vol. 200.

[56] M. Benko, W. Härdle, and A. Kneip, "Common functional principal components," *The Annals of Statistics*, vol. 37, no. 1, pp. 1–34, Feb. 2009, Publisher: Institute of Mathematical Statistics.

[57] E. Paparoditis and T. Sapatinas, "Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data," *Biometrika*, vol. 103, no. 3, pp. 727–733, Sep. 2016.

[58] A. B. Kashlak, S. Myroshnychenko, and S. Spektor, "Analytic Permutation Testing for functional data ANOVA," *Journal of Computational and Graphical Statistics*, vol. 0, no. ja, pp. 1–24, Apr. 2022, Publisher: Taylor & Francis.

[59] A.-M. Staicu, Y. Li, C. M. Crainiceanu, and D. Ruppert, "Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis," *Scandinavian Journal of Statistics*, vol. 41, no. 4, pp. 932–949, 2014.

[60] G.-M. Pomann, A.-M. Staicu, and S. Ghosh, "A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 65, no. 3, pp. 395–414, 2016, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12130.

[61] V. M. Panaretos, D. Kraus, and J. H. Maddocks, "Second-Order Comparison of Gaussian Random Functions and the Geometry of DNA Minicircles," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 670–682, Jun. 2010.

[62] S. Fremdt, J. G. Steinebach, L. Horváth, and P. Kokoszka, "Testing the Equality of Covariance Operators in Functional Samples," *Scandinavian Journal of Statistics*, vol. 40, no. 1, pp. 138–152, 2013, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9469.2012.00796.x.

[63] G. Boente, D. Rodriguez, and M. Sued, "Testing equality between several populations covariance operators," *Annals of the Institute of Statistical Mathematics*, vol. 70, no. 4, pp. 919–950, 2018.

[64] T. W. Anderson, "An introduction to multivariate statistical analysis," Wiley New York, Tech. Rep., 1962.

[65] P. C. O'Brien, "Robust procedures for testing equality of covariance matrices," *Biometrics*, pp. 819–827, 1992.

[66] N. Sugiura and H. Nagao, "Unbiasedness of some test criteria for the equality of one or two covariance matrices," *The Annals of Mathematical Statistics*, vol. 39, no. 5, pp. 1686–1692, 1968.

[67] J. R. Schott, "A test for the equality of covariance matrices when the dimension is large relative to the sample sizes," *Computational Statistics & Data Analysis*, vol. 51, no. 12, pp. 6535–6542, Aug. 2007.

[68] M. S. Srivastava and H. Yanagihara, "Testing the equality of several covariance matrices with fewer observations than the dimension," *Journal of Multivariate Analysis*, vol. 101, no. 6, pp. 1319–1329, Jul. 2010.

[69] T. Cai, W. Liu, and Y. Xia, "Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings," *Journal of the American Statistical Association*, vol. 108, no. 501, pp. 265–277, Mar. 2013, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2012.758041.

[70] M. Cao, T. He, W. Zhou, M. W. Zhou, and T. LazyData, "Package 'hdtest'," *R package version*, 2018.

[71] A Cabassi and A. Kashlak, "Fdcov: Analysis of covariance operators," *R package version*, vol. 1, no. 0, 2016.

[72] H. Hedegaard, Sally C., Curtin, and M. Warner, "Suicide Mortality in the United States, 1999–2017," no. 330, p. 8, 2018.

[73] M. Miller, D. Azrael, and D. Hemenway, "The epidemiology of case fatality rates for suicide in the northeast," *Annals of Emergency Medicine*, vol. 43, no. 6, pp. 723–730, Jun. 2004.

[74] M. Miller, D. Azrael, and C. Barber, "Suicide Mortality in the United States: The Importance of Attending to Method in Understanding Population-Level Disparities in the Burden of Suicide," *Annual Review of Public Health*, vol. 33, no. 1, pp. 393–408, Mar. 2012, Publisher: Annual Reviews.

[75] A. Conner, D. Azrael, and M. Miller, "Suicide Case-Fatality Rates in the United States, 2007 to 2014," *Annals of Internal Medicine*, vol. 171, no. 12, pp. 885–895, Dec. 2019, Publisher: American College of Physicians.

[76] J. M. Bostwick, C. Pabbati, J. R. Geske, and A. J. McKean, "Suicide Attempt as a Risk Factor for Completed Suicide: Even More Lethal Than We Knew," *The American Journal of Psychiatry*, vol. 173, no. 11, pp. 1094–1100, Nov. 2016.

[77] M. D. Anestis, "Prior suicide attempts are less common in suicide decedents who died by firearms relative to those who died by other means," *Journal of Affective Disorders*, vol. 189, pp. 106–109, Jan. 2016.

[78] M. Olfson *et al.*, "National Trends in Suicide Attempts Among Adults in the United States," *JAMA Psychiatry*, vol. 74, no. 11, pp. 1095–1103, Nov. 2017, Publisher: American Medical Association.

[79] J. C. Franklin *et al.*, "Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research," *Psychological Bulletin*, vol. 143, no. 2, pp. 187–232, 2017, Place: US Publisher: American Psychological Association.

[80] B. E. Belsher *et al.*, "Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation," *JAMA psychiatry*, vol. 76, no. 6, pp. 642–651, 2019.

[81] M. Birjali, A. Beni-Hssane, and M. Erritali, "Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks," *Procedia Computer Science*, vol. 113, pp. 65–72, 2017, ISBN: 1877-0509 Publisher: Elsevier.

[82] S. B. Choi, W. Lee, J.-H. Yoon, J.-U. Won, and D. W. Kim, "Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea," *Journal of affective disorders*, vol. 231, pp. 8–14, 2018, ISBN: 0165-0327 Publisher: Elsevier.

[83] J. Torous *et al.*, "Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: A review of current progress and next steps," *Current psychiatry reports*, vol. 20, no. 7, p. 51, 2018, ISBN: 1523-3812 Publisher: Springer.

[84] C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, "Predicting Risk of Suicide Attempts Over Time Through Machine Learning," *Clinical Psychological Science*, vol. 5, no. 3, pp. 457–469, May 2017, Publisher: SAGE Publications Inc.

[85] C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, "Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning," *Journal of Child Psychology and Psychiatry*, vol. 59, no. 12, pp. 1261–1270, 2018.

[86] R. C. Kessler *et al.*, "Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans Health Administration," *International journal of methods in psychiatric research*, vol. 26, no. 3, e1575, 2017, ISBN: 1049-8931 Publisher: Wiley Online Library.

[87] R. C. Kessler, R. M. Bossarte, A. Luedtke, A. M. Zaslavsky, and J. R. Zubizarreta, "Suicide prediction models: A critical review of recent research with recommendations for the

way forward," *Molecular Psychiatry*, vol. 25, no. 1, pp. 168–179, Jan. 2020, Number: 1 Publisher: Nature Publishing Group.

[88] K. Houston, C. Haw, E. Townsend, and K. Hawton, "General practitioner contacts with patients before and after deliberate self harm.," *British Journal of General Practice*, vol. 53, no. 490, pp. 365–370, 2003, ISBN: 0960-1643 Publisher: British Journal of General Practice.

[89] K. Suominen, E. Isometsä, M. Martunnen, A. Ostamo, and J. Lönnqvist, "Health care contacts before and after attempted suicide among adolescent and young adult versus older suicide attempters," *Psychological medicine*, vol. 34, no. 2, p. 313, 2004, ISBN: 0033-2917 Publisher: Cambridge University Press.

[90] B. K. Ahmedani *et al.*, "Health care contacts in the year before suicide death," *Journal of general internal medicine*, vol. 29, no. 6, pp. 870–877, 2014, ISBN: 0884-8734 Publisher: Springer.

[91] J. B. Luoma, C. E. Martin, and J. L. Pearson, "Contact with mental health and primary care providers before suicide: A review of the evidence," *American Journal of Psychiatry*, vol. 159, no. 6, pp. 909–916, 2002, ISBN: 0002-953X Publisher: Am Psychiatric Assoc.

[92] A. Schaffer *et al.*, "Population-based analysis of health care contacts among suicide decedents: Identifying opportunities for more targeted suicide prevention strategies," *World Psychiatry*, vol. 15, no. 2, pp. 135–145, 2016.

[93] B. F. Grant, T. C. Moore, J. Shepard, and K. Kaplan, "Source and accuracy statement: Wave 1 national epidemiologic survey on alcohol and related conditions (NESARC)," *Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism*, vol. 52, 2003.

[94] B. F. Grant, K. K. Kaplan, and F. S. Stinson, "Source and accuracy statement: The wave 2 national epidemiologic survey on alcohol and related conditions," *National Institute on Alcohol Abuse and Alcoholism*, 2007.

[95] Bureau of the Census., "United States Census 2000. Demographic Profiles: 100-percent and Sample Data," Bureau of the Census, Suitland, MD, Tech. Rep., 2007.

[96] D. S. Hasin *et al.*, "The Alcohol Use Disorder and Associated Disabilities Interview Schedule-5 (AUDADIS-5): Procedural validity of substance use disorders modules through clinical re-appraisal in a general population sample," *Drug and alcohol dependence*, vol. 148, pp. 40–46, 2015, ISBN: 0376-8716 Publisher: Elsevier.

[97] G. Canino *et al.*, "The Spanish Alcohol Use Disorder and Associated Disabilities Interview Schedule (AUDADIS): Reliability and concordance with clinical diagnoses in a Hispanic population.," *Journal of studies on alcohol*, vol. 60, no. 6, pp. 790–799, 1999, ISBN: 0096-882X Publisher: Rutgers University Piscataway, NJ.

[98] S. Chatterji, J. B. Saunders, R. Vrasti, B. F. Grant, D. Hasin, and D. Mager, "Reliability of the alcohol and drug modules of the Alcohol Use Disorder and Associated Disabilities Interview Schedule—Alcohol/Drug-Revised (AUDADIS-ADR): An international comparison," *Drug and alcohol dependence*, vol. 47, no. 3, pp. 171–185, 1997, ISBN: 0376-8716 Publisher: Elsevier.

[99] B. F. Grant, D. A. Dawson, F. S. Stinson, P. S. Chou, W. Kay, and R. Pickering, "The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): Reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample," *Drug and Alcohol Dependence*, vol. 71, no. 1, pp. 7–16, Jul. 2003.

[100] D. Hasin, K. M. Carpenter, S. McCloud, M. Smith, and B. F. Grant, "The Alcohol Use Disorder and Associated Disabilities Interview Schedule (AUDADIS): Reliability of alcohol and drug modules in a clinical sample," *Drug Alcohol Depend*, vol. 44, no. 2-3, pp. 133–141, 1997.

[101] B. F. Grant *et al.*, "Sociodemographic and psychopathologic predictors of first incidence of DSM-IV substance use, mood and anxiety disorders: Results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions," *Molecular Psychiatry*, vol. 14, no. 11, pp. 1051–1066, Nov. 2009, Number: 11 Publisher: Nature Publishing Group.

[102] M. Huebner, W. Vach, and S. le Cessie, "A systematic approach to initial data analysis is good research practice," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 151, no. 1, pp. 25–27, Jan. 2016.

[103] R. H. H. Groenwold, I. R. White, A. R. T. Donders, J. R. Carpenter, D. G. Altman, and K. G. M. Moons, "Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis," *CMAJ*, vol. 184, no. 11, pp. 1265–1269, Aug. 2012, Publisher: CMAJ Section: Analysis.

[104] I. R. White and S. G. Thompson, "Adjusting for partially missing baseline measurements in randomized trials," *Statistics in Medicine*, vol. 24, no. 7, pp. 993–1007, 2005, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1981.

[105] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *University of California, Berkeley*, vol. 110, no. 1-12, p. 24, 2004.

[106] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC medical informatics and decision making*, vol. 11, no. 1, p. 51, 2011, ISBN: 1472-6947 Publisher: Springer.

[107] R. A. Poldrack, G. Huckins, and G. Varoquaux, "Establishment of Best Practices for Evidence for Prediction: A Review," *JAMA Psychiatry*, vol. 77, no. 5, pp. 534–540, May 2020, Publisher: American Medical Association.

[108] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[109] M. N. Wright, A. Ziegler, and I. R. König, "Do little interactions get lost in dark random forests?" *BMC Bioinformatics*, vol. 17, no. 1, p. 145, Mar. 2016.

[110] R. C. Kessler, "Clinical Epidemiological Research on Suicide-Related Behaviors—Where We Are and Where We Need to Go," *JAMA Psychiatry*, vol. 76, no. 8, pp. 777–778, Aug. 2019, Publisher: American Medical Association.

[111] J. A. Gordon, S. Avenevoli, and J. L. Pearson, "Suicide prevention research priorities in health care," *JAMA psychiatry*, vol. 77, no. 9, pp. 885–886, 2020.

[112] I. E. Fedyszyn, A. Erlangsen, C. Hjorthøj, T. Madsen, and M. Nordentoft, "Repeated suicide attempts and suicide among individuals with a first emergency department contact for attempted suicide: A prospective, nationwide, Danish register-based study.," *The Journal of clinical psychiatry*, vol. 77, no. 6, pp. 832–840, 2016, ISBN: 0160-6689.

[113] K. Hawton *et al.*, "Suicide following self-harm: Findings from the multicentre study of self-harm in England, 2000–2012," *Journal of Affective Disorders*, vol. 175, pp. 147–151, 2015, ISBN: 0165-0327 Publisher: Elsevier.

[114] C.-J. Kuo, D. Gunnell, C.-C. Chen, P. S. Yip, and Y.-Y. Chen, "Suicide and non-suicide mortality after self-harm in Taipei City, Taiwan," *The British Journal of Psychiatry*, vol. 200, no. 5, pp. 405–411, 2012, ISBN: 0007-1250 Publisher: Cambridge University Press.

[115] J. D. Ribeiro *et al.*, "Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: A meta-analysis of longitudinal studies," *Psychological medicine*, vol. 46, no. 2, pp. 225–236, 2016, ISBN: 0033-2917 Publisher: Cambridge University Press.

[116] J. D. Ribeiro, X. Huang, K. R. Fox, and J. C. Franklin, "Depression and hopelessness as risk factors for suicide ideation, attempts and death: Meta-analysis of longitudinal studies," *The British Journal of Psychiatry*, vol. 212, no. 5, pp. 279–286, 2018, ISBN: 0007-1250 Publisher: Cambridge University Press.

[117] K. Kroenke and R. L. Spitzer, "The PHQ-9: A New Depression Diagnostic and Severity Measure," *Psychiatric Annals*, vol. 32, no. 9, pp. 509–515, Sep. 2002, Publisher: SLACK Incorporated.

[118] K. Posner *et al.*, "Columbia-suicide severity rating scale (C-SSRS)," *New York, NY: Columbia University Medical Center*, vol. 2008, 2008.

[119] A. T. Beck, R. A. Steer, and M. G. Carbin, "Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation," *Clinical Psychology Review*, vol. 8, no. 1, pp. 77–100, Jan. 1988.

[120] W. M. Patterson, H. H. Dohn, J. Bird, and G. A. Patterson, "Evaluation of suicidal patients: The sad persons scale," *Psychosomatics*, vol. 24, no. 4, pp. 343–349, 1983.

[121] A Nimeus, M Alsen, and L. Träskman-Bendz, "The suicide assessment scale: An instrument assessing suicide risk of suicide attempters," *European Psychiatry*, vol. 15, no. 7, pp. 416–423, 2000.

[122] A. T. Beck, "Measuring depression: The depression inventory," *Recent advances in the psychobiology of the depressive illnesses*, pp. 299–302, 1972.

[123] C. A. McHorney, J. E. Ware Jr, J. R. Lu, and C. D. Sherbourne, "The mos 36-item short-form health survey (sf-36): Iii. tests of data quality, scaling assumptions, and reliability across diverse patient groups," *Medical care*, pp. 40–66, 1994.

[124] G. Vilagut *et al.*, "The mental component of the short-form 12 health survey (SF-12) as a measure of depressive disorders in the general population: Results with three alternative scoring methods," *Value in Health*, vol. 16, no. 4, pp. 564–573, 2013, ISBN: 1098-3015 Publisher: Elsevier.

[125] J. E. Ware Jr, M. Kosinski, and S. D. Keller, "A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity," *Medical care*, pp. 220–233, 1996, ISBN: 0025-7079 Publisher: JSTOR.

[126] K. Hawton, "Sex and suicide: Gender differences in suicidal behaviour," *The British Journal of Psychiatry*, vol. 177, no. 6, pp. 484–485, 2000, ISBN: 0007-1250 Publisher: Cambridge University Press.

[127] K. Skogman, M. Alsén, and A. Öjehagen, "Sex differences in risk factors for suicide after attempted suicide," *Social psychiatry and psychiatric epidemiology*, vol. 39, no. 2, pp. 113–120, 2004, ISBN: 1433-9285 Publisher: Springer.

[128] A. Case and A. Deaton, "Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century," *Proceedings of the National Academy of Sciences*, vol. 112, no. 49, pp. 15 078–15 083, 2015, ISBN: 0027-8424 Publisher: National Acad Sciences.

[129] J. A. Kaufman, L. K. Salas-Hernández, K. A. Komro, and M. D. Livingston, "Effects of increased minimum wages by unemployment rate on suicide in the USA," *J Epidemiol Community Health*, vol. 74, no. 3, pp. 219–224, 2020, ISBN: 0143-005X Publisher: BMJ Publishing Group Ltd.

[130] M. A. Reger, I. H. Stanley, and T. E. Joiner, "Suicide Mortality and Coronavirus Disease 2019—A Perfect Storm?" *JAMA Psychiatry*, Apr. 2020.

[131] K. N. Fountoulakis *et al.*, "Rate of suicide and suicide attempts and their relationship to unemployment in Thessaloniki Greece (2000–2012)," *Journal of affective disorders*, vol. 174, pp. 131–136, 2015, ISBN: 0165-0327 Publisher: Elsevier.

[132] A. Reeves, M. McKee, and D. Stuckler, "Economic suicides in the Great Recession in Europe and North America," *The British Journal of Psychiatry*, vol. 205, no. 3, pp. 246–247, Sep. 2014, Publisher: Cambridge University Press.

[133] D. Stuckler, S. Basu, M. Suhrcke, A. Coutts, and M. McKee, "The public health effect of economic crises and alternative policy responses in Europe: An empirical analysis," *The Lancet*, vol. 374, no. 9686, pp. 315–323, Jul. 2009.

[134] H. Meltzer, P. Bebbington, T. Brugha, R. Jenkins, S. McManus, and M. S. Dennis, "Personal debt and suicidal ideation," *Psychological medicine*, vol. 41, no. 4, p. 771, 2011, ISBN: 0033-2917 Publisher: Cambridge University Press.

[135] E. Y. Chen *et al.*, "Suicide in Hong Kong: A case-control psychological autopsy study," *Psychological medicine*, vol. 36, no. 6, p. 815, 2006, ISBN: 0033-2917 Publisher: Cambridge University Press.

[136] E. B. Elbogen, M. Lanier, A. E. Montgomery, S. Strickland, H. R. Wagner, and J. Tsai, "Financial Strain and Suicide Attempts in a Nationally Representative Sample of US Adults," *American Journal of Epidemiology*,

[137] *NOT-OD-20-057: Notice of Information: National Death Index Linkage Access for NIH-Supported Investigators*, Jan. 2020.

[138] A. Z. Ivey-Stephenson, A. E. Crosby, S. P. D. Jack, T. Haileyesus, and M.-j. Kresnow-Sedacca, "Suicide Trends Among and Within Urbanization Levels by Sex, Race/Ethnicity, Age Group, and Mechanism of Death — United States, 2001–2015," *MMWR Surveillance Summaries*, vol. 66, no. 18, pp. 1–16, Oct. 2017.

[139] B. Mars *et al.*, "Using Data Linkage to Investigate Inconsistent Reporting of Self-Harm and Questionnaire Non-Response," *Archives of Suicide Research*, vol. 20, no. 2, pp. 113–141, Apr. 2016, Publisher: Routledge _eprint: https://doi.org/10.1080/13811118.2015.1033121.

[140] M. A. Hom *et al.*, "Investigating the reliability of suicide attempt history reporting across five measures: A study of US military service members at risk of suicide," *Journal of Clinical Psychology*, vol. 75, no. 7, pp. 1332–1349, 2019.

Appendix A: Supplement to Two Sample Hypothesis Testing for Functional Principal Components

A.1 Sensitivity Analysis for tuning K

We present sensitivity analysis to explore the role that choosing the number of Functional Principal Components play in the numerical properties our proposed methodology. Our goal is to compare the methods when fixing K, and the impact K has on the empirical rejection rates. We generate synthetic datasets for two groups of independent functional data as described in Section 3.3.1. For each dataset, we estimate K using three PVE thresholds (95%, 99%, 99.9%) and implement the three methods previously described in Section 3.3.1 with the K's selected from the three criteria.

Figure A.1 summarize the simulation rejection rates for the three PVE thresholds used (95%, 99%, 99.9%). We find that our test is more powerful when choosing a 99.9% for the scenario in which the eigenfunctions are different. We found that both [61] and [60]'s tests perform better for smaller values of K when the FPCs in two groups are the same, but the eigenvalues differ. We found our method to be at least equally powerful as competing methodologies when fixing the value of K.

We now explore the paired data scenario by generating synthetic datasets for two groups of paired functions as described in Section **??**. For each dataset, we estimate K using three PVE thresholds (95%, 99%, 99.9%) and implement the four methods previously described in that same section. Figures A.2 summarize the simulation rejection rates for the three PVE thresholds used (95%, 99%, 99.9%). We found that our test is somehow stable across values of PVE and that both [61] and [60]'s tests perform better for smaller values of K when the FPCs in two groups are the same, but the eigenvalues differ. Our method is at least equally powerful as competing

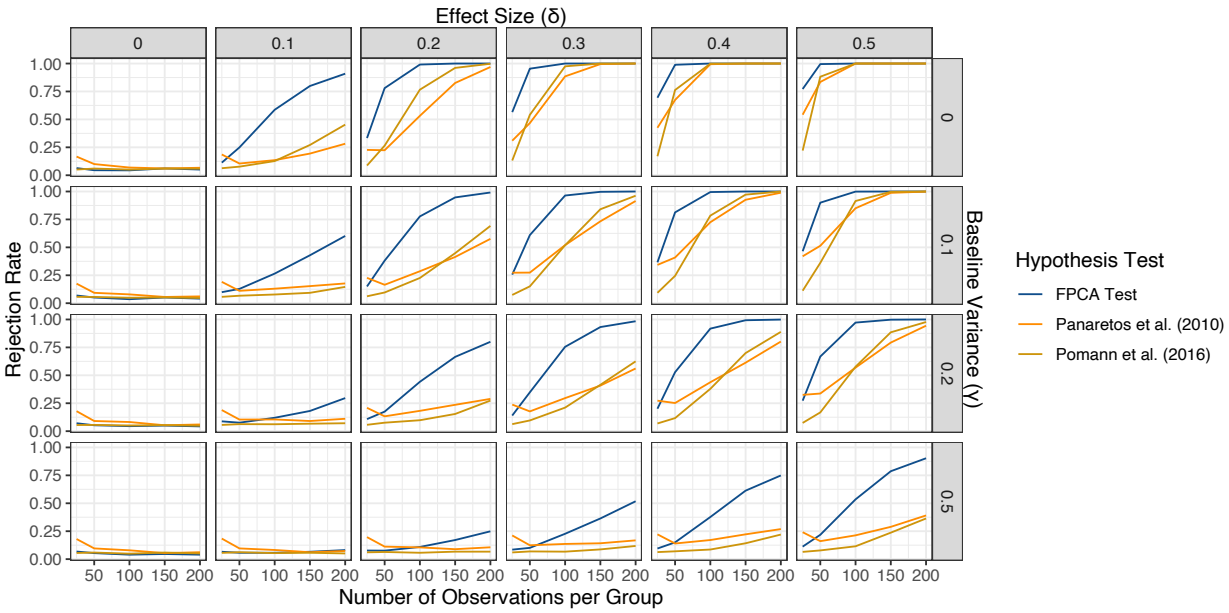


Figure A.1: Plot of Empirical Rejection Rates for Independent Data when PVE = (95%, 99%, 99.9%). We run 1000 simulations for each scenario and reject the null hypothesis at a 5% level. Our proposed test for independent data is in dark blue, [61]'s test is in orange and [60]'s test is in yellow. We present the scenario $\gamma = 0.5$

methodologies when fixing the value of K.

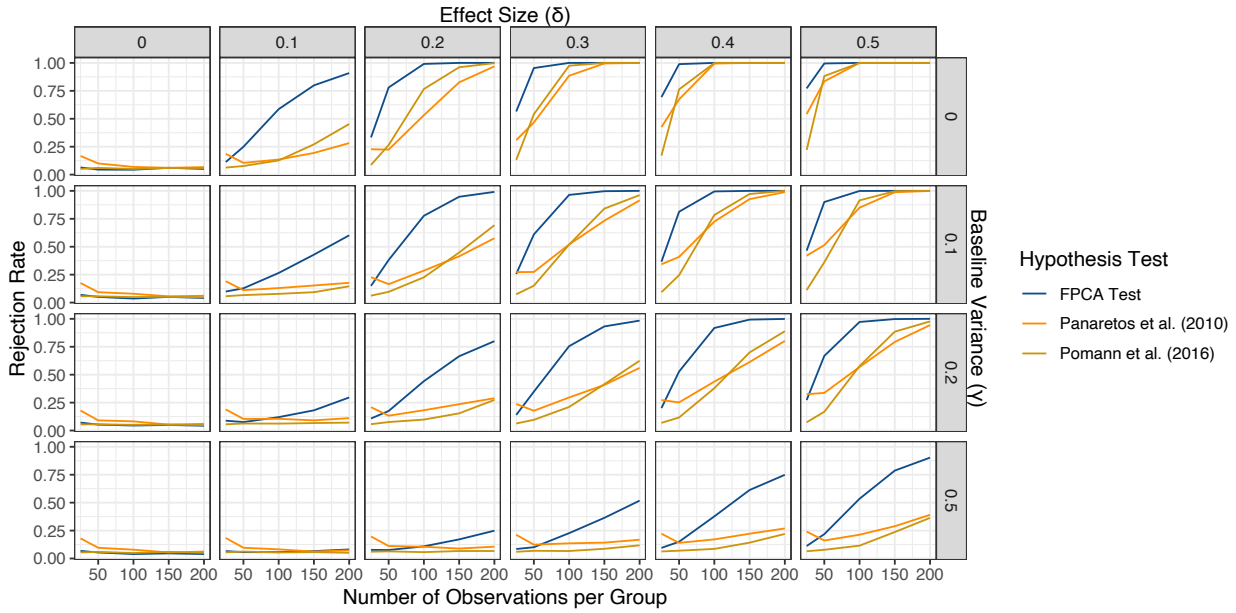


Figure A.2: Plot of Empirical Rejection Rates for Paired Data when PVE = 99%. We run 1000 simulations for each scenario and reject the null hypothesis at a 5% level. We use the percent variance explained criterion and only test the first FPCs that explain 95% of the variance. Our proposed paired data test is in dark blue, and the proposed independent test is in light blue. The test in [61] is in orange and the test in [60] is in yellow. We present the scenario $\gamma = 0.5$ and $\delta = 0.5$

Appendix B: Supplement to Identification of suicide attempt risk factors in a national U.S. survey using machine learning

B.1 Methods Supplement

B.1.1 Predictors from Wave 1

Supplemental eTable B.1 summarizes the response patterns across each of the 14 sections from the Alcohol Use Disorder and Associated Disabilities Interview Schedule DSM-IV version (AUDADIS-IV). To reduce interview burden, participants skipped entire sections based on their responses to gate questions. For example, if participants did not report any alcohol use, they were not asked questions about alcohol use disorders. However, participants could not skip out of other sections, such as sociodemographic characteristics, family history of alcohol and drug use disorders, and family history of major depressive disorder. The survey's total number of questions is 2805, and the minimum number of questions any participant was asked was 474. In all, there were between 643 and 2,985 features available as risk factors for each Wave 1 subject, including 180 DSM-IV derived diagnosis variables.

B.2 Data Analysis

B.2.1 Organizing the predictor variables

Following recommendations [102], we performed an initial data analysis and pre-processed the entire questionnaire by classifying features as categorical (i.e., one category for each response option) unless there was evidence the feature should be considered ordered or continuous. We defined features as ordered or continuous if there were more than 20 response ordered values (e.g., "What is your weight in pounds?") or the label of the question included "number" as part of its

name or description (e.g., "What is the largest number of glasses/containers of wine consumed on days when you drank wine in the last 20 months?"). Once ordered or continuous variables were identified, they were categorized into tertiles. Because gate questions led participants to skip some questions, these informative structural missing values were addressed by creating an indicative new label within each variable. This response category was used to differentiate it from a "true" missing response in which participants did not respond to a question that they were asked. To account for both types of missing data (structural due to skips and true missing response) every variable had two missing categories. We used this missing-indicator method [103, 104] as this approach avoids problems associated with using other techniques such as imputation and feature reduction when there is large scale structural missingness, as the case in most population-based surveys. Of the 2,985 separate features from Wave 1, 7 were completely noninformative because all subjects had the same responses or were colinear with another variable. These 7 features were removed, leaving 2,978 for modeling.

B.2.2 Suicide Attempt Model Tuning

We used balanced random forest (BRF) to build a suicide attempt model taking all 2,978 Wave 1 features to classify dichotomous suicide attempt at Wave 2. The balanced random forest is an ensemble of classifications trees, each fitted to a bootstrap sample of the minority class (i.e., those that attempted suicide) and a random draw of the same size from the majority class (i.e., those who had not attempted). Intuitively, the algorithm builds case-control samples and trains a classification tree on each of them. The final risk score is made by aggregating all the trees of the ensemble. In our analysis, we fixed the number of trees to grow to 4,000 to ensure that, on average, each of our controls was sampled into 25 trees.

To summarize model performance, we calculated a receiver operating characteristic (ROC) curve at various classification thresholds. We selected the optimal tuning parameter (number of variables to draw at each of the nodes) as the ones that lead to the highest average AUC across the ten folds. To assess the Balanced Random Forest variables' importance, we calculated the

decrease in classification ability after any individual variable was permuted across the dataset. In the permutation step, we broke the association between our risk factor and outcome, so that the model cannot use this to model suicide attempts [108, 109]. We took this approach instead of refitting the model without one variable at a time since it substantially reduced the computation power needed to calculate the importance of a variable. We ranked our risk factors by the decrease in classification accuracy.

B.2.3 Suicide Risk Stratification

The suicide risk model provides a continuous risk score for future suicide attempt within three years between surveys. To facilitate interpretation and provide a useful description of the risk across the U.S. population we considered one possible way of categorizing/stratifying the model-calculated risk scores. Our first cut point was based on maximizing the Youden's J statistic (sensitivity + specificity – 1), which is commonly used for risk stratification to balance the tradeoff between sensitivity and specificity. This was used to determine Low risk versus anything higher than Low risk and the cut-point corresponded to a sensitivity=85.3% and Specificity=73.3%. This cut-point placed 73.1% of the U.S. population in the Low risk category. To further categorize the remaining 26.9% of the population not in the Low category, we chose two meaningful population benchmarks: 1) the cut-point corresponding to the top decile of risk across the sample (designating the High Risk group), and 2) the cut-point corresponding to a positive predictive value (PPV) of 10% or above (designating the Very High Risk group).). After sample weighting to the U.S. population, cutting at High or above corresponded to identifying 9.4% of the U.S. population (i.e. 7.5% High and 1.8% Very High) and the PPV of the Very High group was 10.4% (Table 1). We note the somewhat arbitrary choice of these cut-points, but we believe these 4-categories provide a useful descriptive quantification of suicide attempt risk in the U.S.

B.2.4 Model Validation

To further validate our model, we examined how well it was able to categorize risk based on the timing between the waves of the new suicide attempt. Stratifying the suicide attempts by the time they after Wave 1 (i.e. within the first year after interview, between 1-2 years after, 2-3 years after, and between year 3 after but before time of Wave 2 interview) we examined the model-calculated risk categories. To assess the validity of the model across various demographic groups, we stratified the classification performance of the model across sex, age, race (white vs. non-white), and income groups. Specifically, we present the total number of participants, the number of cases, the proportion of correctly identified attempters (i.e. using Medium and above categories to indicate model predicted cases), and the AUC, within each stratum.

B.2.5 Comparison of Original Wave 1 sample with Wave 2 Non-Responders

To investigate the possibility of selection bias related to non-response to the Wave 2 survey, we examined the distribution a Wave 1 demographics and also factors identified to be important for suicide attempt risk by: the total Wave 1 sample (n=43,093), Wave 2 study participants with (n=222) and without (n=34,431) suicide attempts during follow-up, and Wave 2 non-responders (n= 8440) (Supplemental eTable B.3). Those who were lost to follow-up at Wave 2 tended to be older in age (48.15 [21.08] vs. 46.04 [17.38] years), more males (47% vs. 43%), non-white (49% vs. 43%), and lower income (56% vs. 49%). We note that these differences in unweighted demographic characteristics are adjusted by the NESARC Wave 2 non-response weights in all analyses. That is, in the modeling results we incorporated the Wave 2 sampling/non-response weights provided by NESARC which adjust the demographic distribution at Wave 2 to be the same as Wave 1.

Focusing on non-demographic factors related to suicide risk, we found that participants who were Wave 2 non-responders had endorsed suicide behavior symptoms at Wave 1 at a lower rate than Wave 2 responders (both the participants with and without a suicide attempt between waves). Specifically, 2.5% of the full Wave 1 sample endorsed lifetime suicide attempt with depressed

mood before Wave 1, whereas only 2.0% (N=171) of the Wave 2 non-responders endorsed this at Wave 1. Similarly, suicidal ideation was lower at Wave 1 for Wave 2 non-responders (6.3% Wave 2 non-responders thought about committing suicide in lifetime at Wave 1 vs 8.3% of the full sample; 8.5% of Wave 2 non-responders felt like wanted to die vs 10% in the full sample). A similar prevalence of recent financial crisis was found for the full sample at wave 1 versus the Wave 2 non-responders (12% vs 11%).

Table B.1: Summary of response patterns across wave 1 NESARC[a] survey sections.

Section	Number of Questions in Section	Minimum Number of Questions Asked
Section 1: Background Information	117	85
Section 2: Alcohol		
2A: Alcohol Consumption	72	3
2B: Alcohol Abuse	126	0
2C: Alcohol Treatment Utilization	59	0
2D: Family history of Alcoholism	38	38
Section 3: Drug Abuse		
3A: Tobacco use and Dependence	145	5
3B: Drug Use	107	10
3C: Drug Abuse / Experiences	1075	0
3D: Drug Treatment Utilization	61	0
3E: Family History of Drug Abuse	22	22
Section 4: Major Depression (Low Mood)		
4A: Major Depression (Symptoms)	79	2
4B: Family History of MD	22	22
4C: Dysthymia	58	1
Section 5: Mania or Hypomania High Mood	68	3
Section 6: Panic Disorders and Agoraphobia (Anxiety)	82	3
Section 7: Social Phobia	95	3
Section 8: Specific Phobia	85	13
Section 9: Generalized Anxiety	103	2

Section 10: Personality Disorders	110	110
Section 11: Antisocial Personality Disorder		
11A: Antisocial Personality Disorder (Behavior)	117	99
11B: Family History of Antisocial Personality Disorder	22	22
Section 12: Pathological Gambling	107	1
Section 13: Medical Conditions	35	30
Total separate questions	2805	474
NESARC publicly available derived variables created from questions above (e.g. DSM-IV Diagnoses)	180	169
Total features	2985	643

Footnote:
[a] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005.

Table B.2: Summary of the predicted risk scores [a] based on NESARC [b] wave one responses (P = 2985) weighted to be representative of the U.S. population [c] broken down by year of suicide attempt (N = 222)

		Risk Group Proportion [c,d]				
Time to Suicide Attempt	Mean Predicted Score[e]	Low	Medium	High	Very High	N
Within the first year	66.41	8.39%	20.13%	21.02%	50.46%	50
Between the first and second year	58.44	16.29%	20.43%	30.17%	33.11%	52
Between the second and third year	60.45	9.20%	18.96%	41.55%	30.30%	71
Between the third year and wave 2 interview	51.02	23.59%	32.65%	28.08%	16.48%	49

Footnote:
[a] Predictions are based on the out-of-sample prediction from our machine learning balanced random forest model
[b] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005.
[c] Data is weighted to be representative of the U.S. adult population for region, age, sex, race, and ethnicity, based on the 2000 Census
[d] The predicted risk groups are defined on interpretable thresholds for suicide risk prediction.
[e] Predictions are based on the out-of-sample prediction from our machine learning balanced random forest model

Table B.3: Comparison of suicide attempt related characteristics from NESARC[a] Wave 1 across overall NESARC[a] Wave 1 sample and Wave 2 Responders and Non-responders sample

Characteristic at Wave 1	Overall Wave 1	Wave 2 Non-Attempters	Wave 2 Attempters	W2 Non-responders
Total	N = 43,093	N = 34,431[a]	N = 222[a]	N = 8,440
Age Mean (SD)	46.40 (18.18)	46.04 (17.38)	34.98 (12.37)	48.15 (21.08)
Sex				
Female	24,575 (57%)	19,951 (58%)	138 (62%)	4,486 (53%)
Male	18,518 (43%)	14,480 (42%)	84 (38%)	3,954 (47%)
Race				
White	24,507 (57%)	20,055 (58%)	119 (54%)	4,333 (51%)
Non-White	18,586 (43%)	14,376 (42%)	103 (46%)	4,107 (49%)
Personal Income				
$19,999	21,075 (49%)	16,212 (47%)	154 (69%)	4,709 (56%)
$20,000 - $34,999	9,999 (23%)	8,066 (23%)	44 (20%)	1,889 (22%)
$35000-$69999	9,031 (21%)	7,631 (22%)	18 (8.1%)	1,382 (16%)
$69,999	2,988 (6.9%)	2,522 (7.3%)	6 (2.7%)	460 (5.5%)
Attempted Suicide				
Yes	1,074 (2.5%)	823 (2.4%)	80 (36%)	171 (2.0%)
No	12,596 (29%)	10,401 (30%)	72 (32%)	2,123 (25%)
N/A	29,340 (68%)	23,145 (67%)	69 (31%)	6,126 (73%)
Missing	83 (0.2%)	62 (0.2%)	1 (0.5%)	20 (0.2%)
Thought about committing suicide				
Yes	3,566 (8.3%)	2,925 (8.5%)	110 (50%)	531 (6.3%)
No	10,098 (23%)	8,296 (24%)	42 (19%)	1,760 (21%)
N/A	29,340 (68%)	23,145 (67%)	69 (31%)	6,126 (73%)
Missing	89 (0.2%)	65 (0.2%)	1 (0.5%)	23 (0.3%)
Felt like wanted to die				
Yes	4,468 (10%)	3,631 (11%)	122 (55%)	715 (8.5%)
No	9,163 (21%)	7,561 (22%)	30 (14%)	1,572 (19%)
N/A	29,340 (68%)	23,145 (67%)	69 (31%)	6,126 (73%)
Missing	122 (0.3%)	94 (0.3%)	1 (0.5%)	27 (0.3%)
During past 4 weeks, how often felt downhearted and depressed				
All / Most of the time	2,958 (6.9%)	2,201 (6.4%)	61 (27%)	696 (8.2%)
Some / A little / None of the time	39,718 (92%)	32,000 (93%)	160 (72%)	7,558 (90%)
N/A	417 (1.0%)	230 (0.7%)	1 (0.5%)	186 (2.2%)

During past 4 weeks, how often did work or other activities less carefully than usual as result of emotional problems				
All / Most of the time	2,836 (6.6%)	1,949 (5.7%)	49 (22%)	838 (9.9%)
Some / A little / None of the time	39,874 (93%)	32,260 (94%)	173 (78%)	7,441 (88%)
N/A	383 (0.9%)	222 (0.6%)	0 (0%)	161 (1.9%)
Experienced major financial crisis, bankruptcy or been unable to pay bills on time in last 12 months				
Yes	5,037 (12%)	4,000 (12%)	92 (41%)	945 (11%)
Yes	37,679 (87%)	30,223 (88%)	130 (59%)	7,326 (87%)
Unknown	377 (0.9%)	208 (0.6%)	0 (0%)	169 (2.0%)

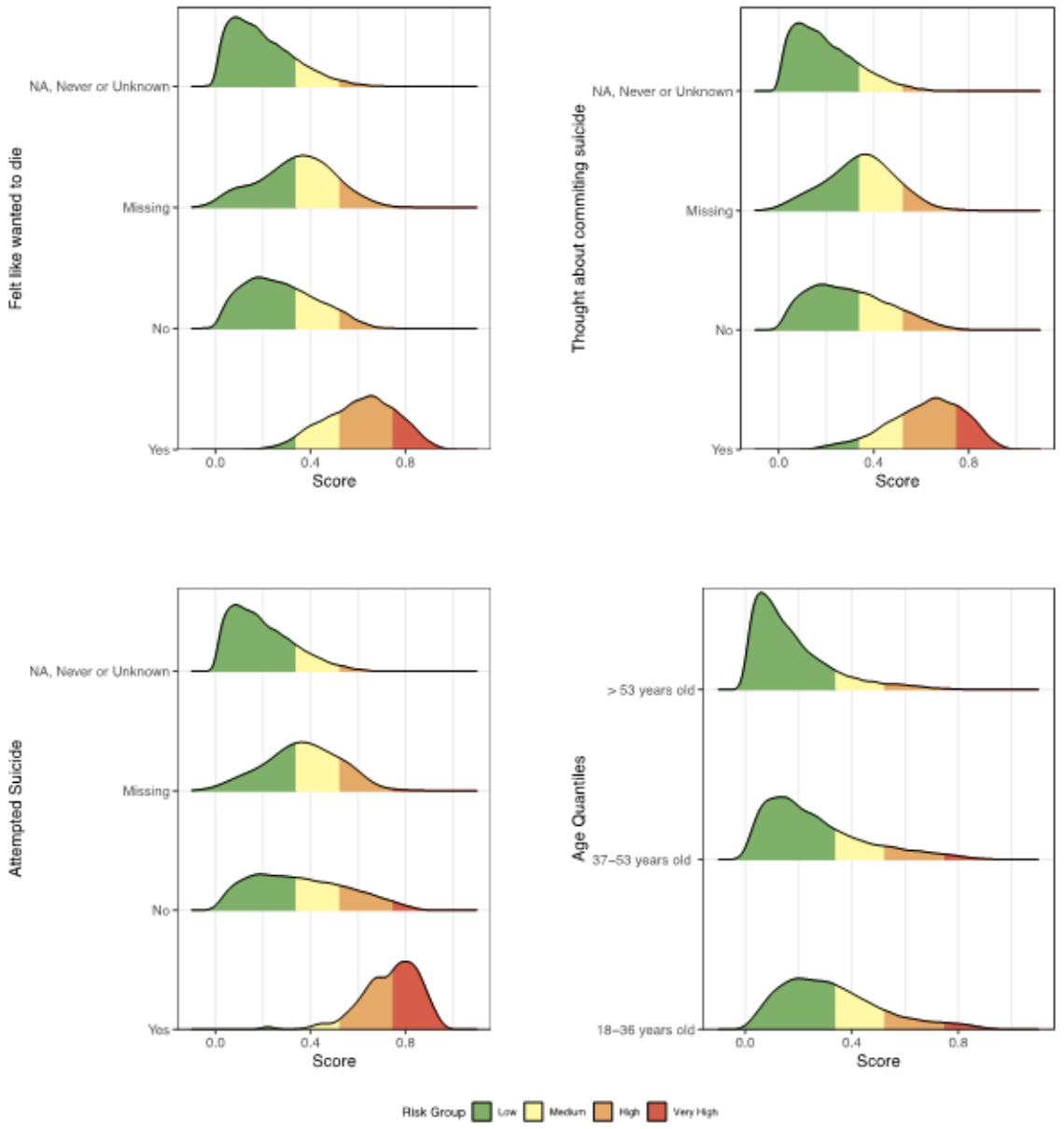
Footnote:
[a] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005.
[b] Statistics presented: mean (SD); n (%)

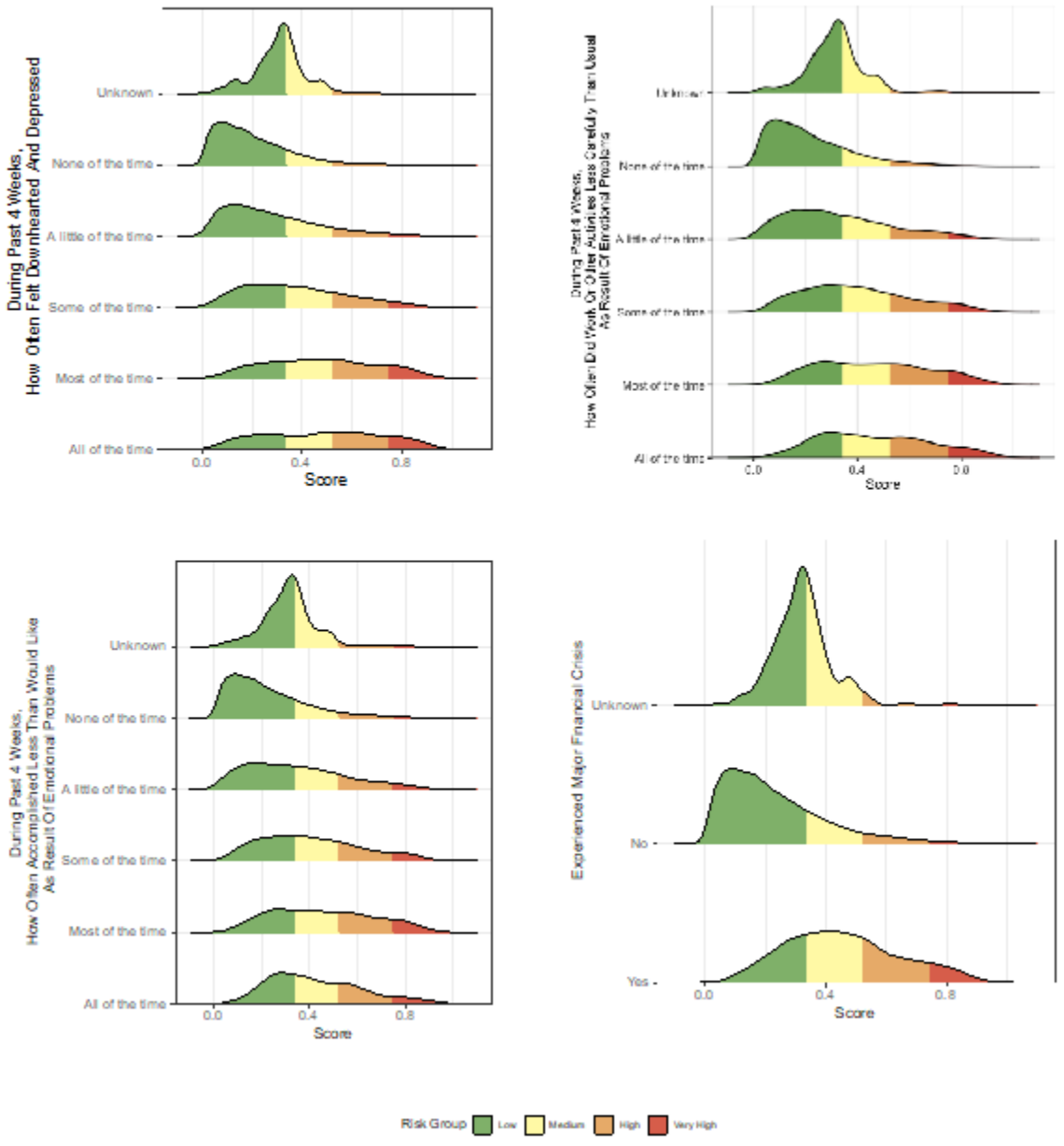
Table B.4: Summary of suicide attempts and model prediction performance [a] based on NESARC [b] wave one demographics

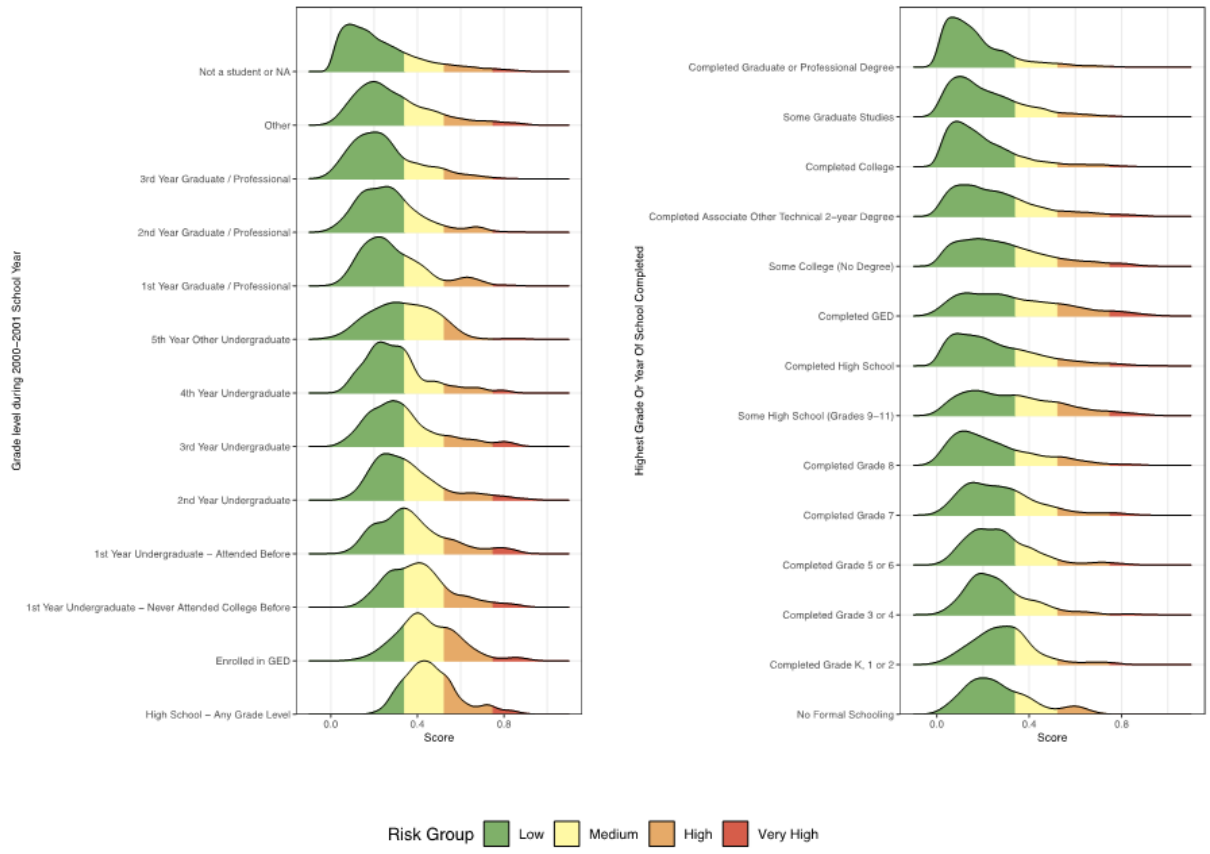
	Total	Proportion of Sample	N Attempted	N Did not attempt	N of Correctly Identified Attempters [c]	Sensitivity [c]	Specificity [c]	AUC (95% Confidence Interval)
Age								
18-36 years old	11551	0.333	129	11422	111	0.860	0.593	0.808 (0.765-0.851)
37-53 years old	11551	0.333	74	11477	65	0.878	0.708	0.867 (0.827-0.906)
>53 years old	11551	0.333	19	11532	14	0.737	0.862	0.872 (0.8-0.945)
Sex								
Female	20089	0.580	138	19951	116	0.841	0.705	0.850 (0.813-0.886)
Male	14564	0.420	84	14480	74	0.881	0.743	0.872 (0.84-0.904)
Race								
White	20174	0.582	119	20055	104	0.874	0.747	0.877 (0.845-0.909)
Non-White	14479	0.418	103	14376	86	0.835	0.684	0.831 (0.788-0.873)
Income								
$0-20000	16366	0.472	154	16212	135	0.877	0.632	0.845 (0.813-0.877)
$20000-34999	8110	0.234	44	8066	38	0.864	0.749	0.848 (0.786-0.91)
$35000-69,999	7649	0.221	18	7631	12	0.667	0.824	0.794 (0.676-0.911)
$70,000 or more	2528	0.073	6	2522	5	0.833	0.893	0.944 (0.893-0.994)

Footnote:
[a] Predictions are based on the out-of-sample prediction from our machine learning balanced random forest model
[b] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005.
[c] We define correctly identified as being predicted in the medium risk group or higher

Figure B.1: Response plots [a] for 10 most important variables from suicide prediction model using NESARC [b] wave one responses (P = 2985) colored by predicted risk group [c]




During Past 4 Weeks, How Often Felt Downhearted And Depressed
Unknown
None of the time
A little of the time
Some of the time
Most of the time
All of the time
0.0
0.4
0.8
Score
During Past 4 Weeks, How Often Did Work Or Other Activities Less Carefully Than Usual As Result Of Emotional Problems
Unknown
None of the time
A little of the time
Some of the time
Most of the time
All of the time
0.0
0.4
0.8
Score
During Past 4 Weeks, How Often Accomplished Less Than Would Like As Result Of Emotional Problems
Unknown
None of the time
A little of the time
Some of the time
Most of the time
All of the time
0.0
0.4
0.8
Score
Experienced Major Financial Crisis
Unknown
No
Yes
0.0
0.4
0.8
Score
Risk Group
Low
Medium
High
Very High



Footnote:

[a] Response plots show the distribution of the predicted risk scores broken down by the responses to each of the individual questions

[b] NESARC: National Epidemiologic Survey on Alcohol and Related Conditions; wave 1 was conducted in 2001 and 2002, and wave 2 in 2004 and 2005

[c] The predicted risk groups are defined on interpretable thresholds for suicide risk prediction.