

Fast Generalized Functional Principal Components Analysis

Andrew Leroux

Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus

Ciprian M. Crainiceanu

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Julia Wrobel

Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus

June 6, 2023

Abstract

We propose a new fast generalized functional principal components analysis (fast-GFPCA) algorithm for dimension reduction of non-Gaussian functional data. The method consists of: (1) binning the data within the functional domain; (2) fitting local random intercept generalized linear mixed models in every bin to obtain the initial estimates of the person-specific functional linear predictors; (3) using fast functional principal component analysis to smooth the linear predictors and obtain their eigenfunctions; and (4) estimating the global model conditional on the eigenfunctions of the linear predictors. An extensive simulation study shows that fast-GFPCA performs as well or better than existing state-of-the-art approaches, it is orders of magnitude faster than existing general purpose GFPCA methods, and scales up well with both the number of observed curves and observations per curve. Methods were motivated by and applied to a study of active/inactive physical activity profiles obtained from wearable accelerometers in the NHANES 2011-2014 study. The method can be implemented by any user familiar with mixed model software, though the R package `fastGFPCA` is provided for convenience.

Keywords: functional data, FPCA, generalized FPCA

1 Introduction

Functional data analysis (FDA) Ramsay and Silverman (2005) provides a wide range of analysis techniques for data with complex structures such as time series or images. These data are often high dimensional (many repeated observations per function) and exhibit complex and non-stationary patterns of variation. Functional principal components analysis (FPCA) Jones and Rice (1992); Rice and Silverman (1991); Staniswalis and Lee (1998), the analog of principal components analysis (PCA) Hotelling (1933); Jolliffe (1982); Pearson (1901) for functional data, is a first-line dimension reduction technique for the analysis for such data. Key differences between FPCA and PCA are that functional data: (1) may be observed with substantial measurement error; (2) is expressed in the same unit of measurement at every point in the domain; and (3) has a functional domain that is ordered and has a natural distance (e.g., time ordering and distance).

There is a rich literature on FPCA focusing on both estimation and inference. Broadly, FPCA methods for Gaussian data involve either smoothing the covariance function Staniswalis and Lee (1998); Xiao et al. (2016); Yao et al. (2003) or estimation via model-based approaches with explicit likelihood assumptions van der Linde (2008); James et al. (2000). Extensions to sparse or irregularly observed data J. and Paul (2009); Yao et al. (2005); Xiao et al. (2018), multivariate Happ and Greven (2015); Chiou et al. (2014); Li and Xiao (2021) and multilevel functional Di et al. (2009); Cui et al. (2022) data exist. Although the FPCA literature is quite extensive, there are few high quality, open source, software implementations. The covariance smoothing FACE approach of Xiao et al. (2021, 2016), implemented in the `refund` package Goldsmith et al. (2020) in R, is by far the fastest approach for estimating FPCA for regularly observed data. The likelihood-based methods require substantially longer computation times for large data.

Here we focus on extensions of FPCA methods to non-Gaussian outcomes (e.g., binary or count data), which we refer to as Generalized FPCA (GFPICA). More precisely, we focus

on methods that decompose the variability of the person-specific latent functional linear predictors along their main directions of variation. In contrast to the relatively large number of published papers on FPCA, there are far fewer GFPCA papers. A few exceptions are Hall et al. (2008); van der Linde (2009); Gertheiss et al. (2017); Weishampel et al. (2023) for single-level and Chen et al. (2013); Goldsmith et al. (2015) for multilevel GFPCA. Unfortunately, the software implementation problem is even more acute for GFPCA compared to FPCA. Indeed, most published methods either lack accompanying software or are extremely slow for large data. Moreover, current methods require pre-specifying both the number of principal components and basis functions used to estimate the principal components. Assessing sensitivity to these key input parameters is critical when applying GFPCA to a new dataset.

The ever increasing number of studies that collect non-Gaussian functional data of increasing size and complexity require methods that are fast and scalable. Consider, for example, the minute level physical activity data obtained from wearable accelerometers deployed in large epidemiologic studies, such as the National Health and Nutrition Survey (NHANES) and UK Biobank Doherty et al. (2017). In many applications one is interested in the pattern of being active (coded as 1) or inactive (coded as 0) at every minute of the day. Thus, the data at the study participant level is a function observed at 1440 minutes (number of minutes in a day) where the value of the function is either active or inactive at every time point. NHANES contains such data for tens of thousands of study participants, while the UK Biobank for close to 100,000 study participants. Our goal is to provide a scalable GFPCA method that aids in interpretation and analysis of large-scale non-Gaussian functional data from the NHANES accelerometry study by extracting orthogonal directions of variation in the linear predictor space of these functions.

Very recently, Weishampel et al. (2023) proposed an approach which allows for estimation of GFPCA for general exponential family outcomes. Their approach is fast, could be viewed as an alternative, but differs in key ways from fast-GFPCA. Importantly, in our

data application, the approach of Weishampel et al. (2023) has the potential to lead to infinite bias in estimated latent functions. We discuss this point further in Section 2.5, but do not focus on their approach because it was not appropriate for our data application. Aside from Weishampel et al. (2023), general purpose software for GFPCA that accommodates multiple types of exponential family data are slow. However, there is also good news about specific types of outcomes. For example, Wrobel et al. (2019) developed a very fast and efficient binary FPCA procedure that uses an EM algorithm to optimize a variational approximation to the binomial FPCA likelihood. The paper is accompanied by the `registr` package Wrobel (2018); Wrobel and Bauer (2021). To our knowledge this is the only publicly available GFPCA implementation that is fast and viable for large datasets, and thus we compare our fast-GFPCA approach to the `registr::bfpca()` function. In addition, for simulation scenarios with smaller data sets, we compare fast-GFPCA to the two-step GFPCA implementation described in Gertheiss et al. (2017), which is available in the `registr` package as the `registr::gfpca_twoStep()` function. This two-step approach can be used to model several exponential family outcomes but is prohibitively slow for large datasets. The current work adds substantially to the literature by providing a general approach to GFPCA which is readily generalizable to functional regression models of interest for which there are currently no fast implementations for large scale data.

Specifically, we propose a new method, fast-GFPCA, with an entirely different philosophy and implementation strategy than most other approaches. The advantages of fast-GFPCA are that: (1) it can be used for any type of non-Gaussian outcome, not just binary; (2) it is orders of magnitude faster than most other all-purpose GFPCA approaches; (3) the method readily handles missing data; (4) it can easily be extended to account for covariates; and (5) it can be generalized to multilevel, longitudinal or structured functional data. We will briefly discuss these extensions, but leave the details for future work.

The remainder of this manuscript is organized as follows. In Section 2 we present the fast GFPCA (fast-GFPCA) approach. Next, in Section 3, we apply the fast-GFPCA to

active/inactive profiles obtained from wearable accelerometers using data from the National Health and Nutrition Survey (NHANES) 2011-2014 waves. We then illustrate the utility of the fast-GFPCA approach in a simulation study in Section 4. We conclude with a discussion in Section 5.

2 Methods

The observed data structure is of the type $\{s_j, Z_i(s_j)\}$, where $Z_i(s_j)$ is a non-Gaussian functional observation for subject $i \in 1, \dots, N$ at the point $s_j \in S$ for $j \in 1, \dots, J$. We assume that these $\{s_j, Z_i(s_j)\}$ pairs are discrete realizations from a continuous process $\{Z_i(s) : s \in S\}$ such that: (1) $g[E\{Z_i(s)\}] = \beta(s) + b_i(s)$, where $g(\cdot)$ is an appropriate link function, $\beta(s)$ is the population mean function in the linear predictor space, and $b_i(s)$ is the individual deviation from the population mean in the linear predictor space; (2) $b_i(s) \sim \text{GP}(0, K_b)$ is a zero mean Gaussian process with covariance operator K_b . Our goals are to decompose the variability of the latent process $b_i(s)$ along its main directions of variation (obtain the FPCA decomposition) and estimate $b_i(s)$ conditional on these directions of variation. Even though $b_i(s)$ are not directly observed, we can use the Karhunen-Loèvè (KL) expansion $b_i(s) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k(s)$ where $\phi_k : S \rightarrow \mathbb{R}$ are orthonormal eigenfunctions such that $\int_S \phi_k^2(s) ds = \lambda_k$, $\lambda_1 \geq \lambda_2 \geq \dots$ are the eigenvalues, and $\xi_{ik} \sim N(0, \lambda_k)$ are mutually independent subject-specific scores over study participants, i , and eigenvalues, k . Together this leads to the GFPCA model,

$$g(E[Z_i(s)]) = \eta_i(s) = \beta_0(s) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(s). \quad (1)$$

This is very similar to the classical FPCA model with the exception that the noise is not necessarily Gaussian.

2.1 Fast GFPCA

We propose the following fast-GFPCA algorithm to conduct FPCA on latent processes when observed data are non-Gaussian: (1) bin the data along the observed functional domain $\mathcal{S} = 1, \dots, J$ into L bins which may be overlapping; (2) estimate separate local GLMMs with a random intercept in each bin to obtain subject-specific estimates on the linear predictor scale; (3) estimate FPCA on the subject-specific estimates obtained from step (2); (4) re-estimate GFPCA using the estimated eigenfunctions obtained from step (3) at the resolution of the original data. Each of these steps is straightforward, though we provide the R package `fastGFPCA` for convenience, described in Section 2.3. Details on each step are provided in Section 2.2 below.

2.2 Estimation Algorithm

We now provide more details on each of the four steps of the fast-GFPCA algorithm, while the algorithmic presentation is provided in Section S-1 of the supplementary material.

Step 1: Choose how to bin the data. The choice of the number of bins (L) and the bin widths (w_l) will be informed by identifiability and assumed complexity of the underlying latent process. Specifically, suppose we choose L bins, where m_l is the midpoint of the $l = 1, \dots, L$ bin. We use symmetric bins, except on the boundary of the domain S . For data observed on a regularly spaced grid over the domain, the l^{th} bin contains the data at domain values $\mathcal{S}_l = \{s_{m_l-w_l/2}, \dots, s_{m_l}, \dots, s_{m_l+w_l/2}\}$, resulting in $w_l + 1$ points. The binned data are then $[\{Z_i(s_j), l\}, 1 \leq i \leq N, j \in \mathcal{S}_l, 1 \leq l \leq L]$.

If the data are cyclic, as in our data application, the bins may cross the boundary. For example, with minute level accelerometry data, if we let $m_1 = 1$ (activity 00:00-00:01) and $w_l = 6$, then $\mathcal{S}_1 = \{1438, 1439, 1440, 1, 2, 3, 4\}$ (activity 23:57-00:05). When the data are non-cyclic we recommend constructing bins as $\mathcal{S}_l = \{\min(s_{m_l-w_l/2}, s_1), \dots, s_{m_l}, \dots, \max(s_{m_l+w_l/2}, s_J)\}$, resulting in bins with as few as $w_l/2 + 1$ data points.

Step 2: Fit a local GLMM in every bin. Specifically, in each bin $l = 1, \dots, L$ we estimate separate models of the form $g[E\{Z_i(s_j)\}] = \beta_0(s_{m_l}) + b_i(s_{m_l}) = \eta_i(s_{m_l})$ for $s_j \in S_l$. Here $\beta_0(s_{m_l})$ is a fixed effect mean and $b_i(s_{m_l})$ is a random intercept evaluated at the center of the bin, s_{m_l} . From these models we obtain estimates of the global mean, $\widehat{\beta}_0(s_{m_l})$, and predictions of the subject-specific random effects $\widehat{b}_i(s_{m_l})$. These predictions are *not* on the original grid the functions were observed on, but rather the midpoints $\{s_{m_1}, \dots, s_{m_L}\} \subset \mathcal{S}$. Importantly, this model is misspecified because it assumes a constant effect in each bin while, in reality, both $\beta_0(\cdot)$ and $b_i(\cdot)$ vary smoothly over \mathcal{S} . This can lead to biased estimates for $\beta_0(s_{m_l})$, $b_i(s_{m_l})$, which has the potential to induce bias in our estimator for K_b . This bias, under reasonable choice of bin width, is largely absorbed in the eigenvalues of the estimated covariance operator, with eigenfunctions being well estimated. As the method hinges on estimating the eigenfunctions well in step 3, but not necessarily the eigenvalues, this is not a problem for the method. Below, we deviate briefly from the description of the fast-GFPCA algorithm to discuss the issue of bias in more detail as we believe such discussion provides critical insights into why the proposed method works in practice and is reasonable.

Discussion of Binning Induced Bias in Estimation of K_b We argue that the binning procedure induces bias in the estimated latent functions evaluated at the midpoints of each bin which in turn induces bias in the estimated covariance function K_b . The point is most readily shown when we assume that the distribution of observed data are continuous uniform along the domain. For ease of presentation, assume that the domain is $S = [0, 1]$. It follows that the distribution of points follows density $f_s(s) = 1$, $s \in [0, 1]$. Further, conditional on $s \in S_l$ (s is in the l^{th} bin), the density of points is iid uniform with conditional density $f_{s|l}(s) = |S_l|^{-1}$ (inverse of the interval width).

Let the superscript ^{bin} notation denote the estimand under the misspecified model to differentiate between the “true” latent process. The misspecified model in Step 2 estimates

$\eta_i^{\text{bin}}(s_{m_l}) = E[\beta_0(s) + b_i(s)|\{\xi_{ik} : 1 \leq k \leq K\}, s \in \mathcal{S}_l] = E[\beta_0(s)|s \in \mathcal{S}_l] + E[b_i(s)|\{\xi_{ik} : 1 \leq k \leq K\}, s \in \mathcal{S}_l]$, with bias which can be split into population Bias $[\beta_0^{\text{bin}}(s_{m_l})] = \beta_0(s_{m_l}) - E[\beta_0(s)|s \in \mathcal{S}_l]$ and subject-specific Bias $[b_i^{\text{bin}}(s_{m_l})] = b_i(s_{m_l}) - E[b_i(s)|\xi_{ik}, s \in \mathcal{S}_l]$ components. The key to fast-GFPCA is the ability to obtain reasonable estimates for the covariance operator $K_b = \text{Var}[b_i(s)]$ in Step 3. The additive bias in $\beta_0^{\text{bin}}(s_{m_l})$ does not affect the estimator of the covariance operator in Step 3 as the data (i.e. $\text{Cov}(a + X, Y) = \text{Cov}(X, Y)$), so we focus on the effect of the subject specific bias, given by

$$\begin{aligned}\text{Bias } [b_i^{\text{bin}}(s_{m_l})] &= b_i(s_{m_l}) - E[b_i(s)|\xi_{ik}, s \in \mathcal{S}_l] \\ &= \sum_{k=1}^K \phi_k(s_{m_l}) \xi_{ik} - E\left[\sum_{k=1}^K \phi_k(s) \xi_{ik} | \xi_{ik}, s \in \mathcal{S}_l\right] \\ &= \sum_{k=1}^K \phi_k(s_{m_l}) \xi_{ik} - \sum_{k=1}^K \xi_{ik} \int_{s \in \mathcal{S}_l} \phi_k(s) f_{s|l}(s) ds \\ &= \sum_{k=1}^K \phi_k(s_{m_l}) \xi_{ik} - \sum_{k=1}^K \xi_{ik} |S_l|^{-1} \int_{s \in \mathcal{S}_l} \phi_k(s) ds\end{aligned}$$

with $|S_l|^{-1}$ operating as a normalizing constant so that $\lim_{w_l \rightarrow 0} |S_l|^{-1} \int_{s \in \mathcal{S}_l} \phi_k(s) ds = \phi_k(s_{m_l})$. That is, as bin width tends to 0 we recover the eigenfunction evaluated at the midpoint of the current bin. Then

$$\begin{aligned}\text{Cov } [b_i^{\text{bin}}(s_{m_u}), b_i^{\text{bin}}(s_{m_v})] &= \sum_{k=1}^K \lambda_k \left[|S_u|^{-1} \int_{\mathcal{S}_u} \phi_k(s) ds \right] \left[|S_v|^{-1} \int_{\mathcal{S}_v} \phi_k(s) ds \right] \\ &\approx \sum_{k=1}^K \{ \lambda_k |S_u|^{-1} |S_v|^{-1} \} \phi_k(s_{m_u}) \phi_k(s_{m_v}) .\end{aligned}$$

The last approximation is related to the mean value theorem, which states that $\frac{1}{|S_u|} \int_{S_u} \phi_k(s) ds \approx \phi_k(s_{m_u})$ and $\frac{1}{|S_v|} \int_{S_v} \phi_k(s) ds \approx \phi_k(s_{m_v})$ so long as $|S_u|, |S_v|$ are not large relative to the curvature in ϕ_k . If the function $\phi_k(\cdot)$ is constant or linear in any of these intervals, the approximation is actually an equality. This equation shows that by diagonalizing $\text{Cov } [b_i^{\text{bin}}(s_{m_u}), b_i^{\text{bin}}(s_{m_v})]$ we obtain close approximations of the eigenvectors $\phi_k(\cdot)$, but not

of the eigenvalues λ_k . Indeed, the eigenvalues are re-scaled by $|S_u||S_v|$, which is the square of the length of the approximating interval when using constant bin width. This also explains why all the bias in the covariance estimation is absorbed by the eigenvalues and passed on, implicitly, to the scores. Therefore, if we diagonalize the linear predictors obtained in this step we obtain unbiased eigenfunctions but biased covariance, eigenvectors, scores, and subject-specific trajectories.

Step 3: Use the predicted responses on the linear predictor scale $[\{\hat{\eta}_i(s_j), l\}, 1 \leq i \leq N, j \in \mathcal{S}_l]$ to estimate $\hat{K}_b(u, v) = \text{Cov}\{\hat{\eta}_i(u), \hat{\eta}_i(v)\}$ for $u, v \in \{s_{w_1}, \dots, s_{w_L}\}$ via the fast covariance estimation (FACE) method implemented in the `refund::fPCA.face()` function. Obtain the estimated eigenfunctions, $[\{\hat{\phi}_k(s_{w_l})\}, 1 \leq l \leq L, 1 \leq k \leq K]$ of the covariance operator \hat{K}_b where the number of eigenfunctions, K , is selected using, for example, the percent variance explained (e.g., 95%, 99%).

Step 4: Estimating the GFPCA model conditional on the basis functions obtained in Step 3. Because the basis functions are estimated on a different grid from the one of the observed data, we first project each eigenfunction on the rich B-spline basis used in the FACE component of the algorithm in Step 3. This provides an estimate of the eigenfunctions at every point where data was originally sampled. After these projections, fast-GFPCA becomes the following generalized linear mixed model (GLMM)

$$g(E[Z_i(s)] | \{\hat{\phi}_k(s) : 1 \leq s \leq J, 1 \leq k \leq K\}) = \beta_0(s) + \sum_{k=1}^K \xi_{ik} \hat{\phi}_k(s) , \quad (2)$$

where $\xi_{ik} \sim N(0, \sigma_k^2)$ are mutually independent and $\beta_0(s)$ is an unspecified smooth function. That is, given the estimates of $\hat{\phi}_k(\cdot)$, this is a GLMM with K uncorrelated random slopes. Using a principal components decomposition with uncorrelated random slopes simplifies the random effects covariance structure such that only K variance parameters need to be estimated, which contributes to computational efficiency.

If we assume a parametric form for $\beta_0(s)$, any generalized linear mixed model software can be used. For the case when $\beta_0(s)$ is modeled nonparametrically, we estimate Model 2 using the `mgcv::bam()` function Wood et al. (2017) with fast REML smoothing parameter selection and the argument `discrete=TRUE` Li and Wood (2020). This approach is highly computation efficient; see example code in Section 2.3. Alternative software for estimating additive generalized linear mixed models are available and may be faster and more memory efficient in certain situations; specifically, the `mgcv::gamm()` Wood (2011) and `gamm4::gamm4()` Wood et al. (2013) functions, which provide interfaces between the `mgcv` and `nlme` Pinheiro and Bates (2000) and `lme4` Bates et al. (2015) packages, respectively.

2.3 Fast GFPCA: Example Code

A key appeal of fast-GFPCA is that it may be implemented by anyone familiar with mixed model software. The code below illustrates how fast-GFPCA can be implemented on a subset of the binary NHANES active/inactive profiles data used in the application described in Section 3. For illustration we use overlapping windows and bin width $w_l = 10$ for $l = 1, \dots, 1440$ (both w_l and l are expressed in minutes). The general organization of the code follows the four steps of the fast-GFPCA algorithm. This code can also be run using a one-line implementation with the accompanying `fastGFPCA` package.

```
library("refund"); library("tidyverse"); library("lme4")
df      <- read.rds("NHANES_example.rds") # read in the data
J       <- length(unique(df$index))        # dimension

## Step 1: Binning decisions
bin_len <- 10                                # bin width (w)
s_m     <- 1:J                                 # bin midpoints s_{m-1}

## Step 2: Fit local binned GLMMs
fit_ls  <- vector(mode="list", length=J) # empty list to store results
for(j in s_m){                                # loop over bins
    # get indices associated with current bin S_l
    # Note: this data is cyclic, so we look across the domain
    #       using the modulo (%%) function
```

```

sind_j <- (j-bin_len_w/2):(j+bin_len_w/2) %% 1440
sind_j[sind_j == 0] <- 1440
# subset to the current indices
df_j <- df %>%
  filter(index %in% sind[sind_j])
# fit the local, binned GLMM:
#   g(E[Z_i(s)]|s_j \in S_1) = \beta_0(s_{m_1}) + b_i(s_{m_1})
fit_j <- glmer(value ~ 1 + (1|id), data=df_j, family=binomial)
# store results (id, \hat{\eta}_i(s_{m_1}), s_{m_1})
fit_ls[[j]] <- data.frame("id" = 1:N,
                           "eta_i" = coef(fit_j)$id[[1]],
                           "s_m" = j)
}
## bind elements of the list row-wise
fit_df <- bind_rows(fit_ls)

## Step 3: FPCA on the binned latent estimates
f pca_latent <- f pca.face(matrix(fit_df$eta_i, N, J, byrow=FALSE),
                           pve=0.95, argvals=sind, knots=20, lower=0)

## Step 4: Re-fit the GLMM on the full data
##           Note that here we do not need to interpolate the eigenfunctions
##           as the bin midpoints contained the original observations points
# data frame of eigenfunctions (take the first four)
phi_mat <- data.frame("index" = s_m, f pca_latent$efunctions[,1:4])
colnames(phi_mat)[2:5] <- paste0("Phi", 1:4)
# merge the data
df <- df %>% left_join(phi_mat, by="index") %>% mutate(id_fac=factor(id))
# re-fit the model using mgcv::bam()
fit <- bam(value ~ s(index,bs="cc",k=20) +
            s(id_fac,by=Phi1,bs="re") + s(id_fac,by=Phi2,bs="re") +
            s(id_fac,by=Phi3,bs="re") + s(id_fac,by=Phi4,bs="re"),
            data=df, family=binomial, discrete=TRUE, method="fREML")

```

For presentation simplicity, the code assumes that data are ordered by subject and then by the domain of the function. If the data are not sorted like this, slight modifications are needed in Steps 3 and 4 or data can be re-ordered before the code is run. Moreover, the subject identifier needs to be a factor variable to fit the correct model, which is why the variable `id_fac` was created in Step 4. Users unfamiliar with the `mgcv` package may be confused by the syntax in the call to `mgcv::bam()` in Step 4. The expression `s(id_fac, by=Phi1, bs="re")` constructs independent normally distributed random slopes. The same random

effects specification in `lme4` would be

```
fit <- glmer(value~1 + (0+Phi1|id) + (0+Phi2|id) + (0+Phi3|id)+(0+Phi4|id) ,  
            data=df, family=binomial),
```

however, `lme4::glmer()` cannot model $\beta_0(t)$ nonparametrically and a parametric form for $\beta_0(s)$ would need to be specified. We hope that seeing the complete code will: (1) show how easy it is to implement the proposed methods; (2) provide a modular and modifiable platform that can be used in similar situations which require specific adjustments; and (3) support the philosophy that analytic methods are not really methods without supporting software.

For user convenience we have developed `fastGPCA`, an R package available on GitHub at <https://github.com/julia-wrobel/fastGPCA>, which wraps the code for implementing fast-GPCA. The detailed description of the package functionality, including examples for both binomial and Poisson distributed functional data, is contained in the `fastGPCA` vignette associated with the package. Briefly, the primary function in the package is `fast_gfpc()`. Function arguments `overlap` and `binwidth` allow the user to select a bin width and choose whether or not to construct overlapping windows for their data. The argument `family` is fed to the functions `lme4::glmer()` and `mvcv::bam()` and is used to specify an exponential family distribution and link function using the syntax, for example, `family = binomial(link = "logit")`. The `fast_gfpc()` function returns an object of class "`fpc`" and results can be easily visualized using the `refund.shiny` package Wrobel et al. (2016).

2.4 Practical Considerations

Several practical considerations must be taken into account when applying the fast-GPCA method. We discuss these considerations in detail below.

The Need for Step 4. The need for Step 4 stems from the fact that Step 3 produces unbiased estimators of the eigenfunctions but biased estimators of the subject-specific random effects. This is driven primarily by bias in the estimated eigenvalues of the covariance operator, which are then passed onto the scores. This effect can be seen in Figure 1, which plots the estimated scores from step 3 and step 4 versus the true scores (Figure 1(A)), the estimated scores from step 3 versus those from step 4 (Figure 1(B)), and the estimated curves on the log odds scale using step 3 versus step 4. The data are generated as binary functional data according to our simulation study (See Section 4 for details), and fast-GFPCA estimates the true eigenfunctions well in this data. From Figure 1(A) and Figure 1(B), we see that not only are the scores obtained from step 3 biased (far away from the identity line, Figure 1(A)), but that the scores in step 4 are effectively a linear re-scaling of those from step 3 (near perfect correlation, but points away from the identity line, Figure 1(B)). This bias results in estimated log-odds which are substantially less correlated with the true latent functions (Figure 1(C)). Note that the slopes in Figure 1(B) are nearly identical which matches the result shown in Section 2 Step 2 (discussion on binning bias) that the eigenvalues are biased by a constant multiplicative factor (and thus the scores are as well).

Identifiability. Local GLMMs may be non-identifiable or model fitting may not converge. An example is when data are binary and all (or nearly all) observations are either 0 or 1 for every study participant in a particular bin. In this case the local model is non-identifiable. Potential solutions include choosing a wider bin width, imposing a lower or upper bound on the linear predictor scale, or modifying the data. For example, Agresti and Caffo (2000) showed that adding two successes and two failures improves the performance of confidence intervals when estimating a probability from binary data. The idea can be used in our context by adding two successes and two failures in every bin. For our application and simulations, increasing the bin size was enough to ensure excellent performance

of the methods.

Bin width. As mentioned in Section 2, the choice of bin width is an important tuning parameter in the fast GFPCA algorithm. Our simulation study presented in Section 4 illustrates this point. There is a need to balance choosing a bin size that is small enough to estimate the curvature of the latent process but not too small to make GLMM fitting unstable. A possible approach is to consider bins of increased sizes to the level where GLMMs can fit the data. The bin sizes can be increased and the stability of estimators compared. An alternative would be to conduct smoothing at the study specific level and inspect the plots to explore the complexity of the underlying functions.

Non-overlapping versus overlapping windows. The choice of whether to use non-overlapping versus overlapping windows will vary by application. Non-overlapping windows reduces the number of local GLMMs that need to be estimated, though computational gains are minor. In contrast, overlapping windows allows estimation at a finer grid of points s_{m_1}, \dots, s_{m_L} , but may induce spurious correlation in the estimated latent processes. The effects of these auto correlations, if any, are not currently known in applications, though we illustrate that the method can result in under-smoothing in our data application. A possible technical solution could be to adapt the smoothing selection criteria of FACE to allow for auto correlated data, though this exceeds the scope of our current work.

2.5 Alternative Approaches

There are at least two alternatives to the proposed approach based on first smoothing individual curves. First, one may consider first smoothing the binary functional data on the response scale and ignoring the binary nature of the data. These smoothed data may then be transformed using a link function (e.g., logit for binary data), and then apply fPCA to the estimates on the latent scale. Alternatively, smoothing may be done by fitting

generalized additive models $g(E[Y_i(s)]) = \beta_{0i} + f_i(s)$ separately for each individual, allowing the linear predictor to vary smoothly over the domain where $f_i(s)$ is modelled using a rich basis. fPCA may then be applied to the resulting estimates on the latent scale. The latter approach was recently proposed by Weishampel et al. (2023). Both of these approaches are intuitive and may be faster in certain scenarios. While they could be used as quick exploratory tools, we will point out some of their hidden drawbacks.

Consider the first approach. If smoothing is done on the response scale and we denote the smoothed response as $\tilde{Y}_i(s)$, a fundamental requirement is for $g\{\tilde{Y}_i(s)\}$ to be finite and defined. For example, for binary data one would need $\tilde{Y}_i(s) \in (0, 1)$ and preferably farther away from the boundaries to ensure that $\text{logit}\{\Pr(\tilde{Y}_i(s) = 1)\}$ is defined and well behaved. In areas of where there are many zeros (e.g., during the night for physical activity) and ones (during the morning for physical activity), this smooth estimators will be 0 or 1, respectively. Thus, the smooth estimates would need to be bounded away from 0 and 1 using artificial tuning parameters to inverse transform them. This makes the approach dependent on individual choices and may introduce infinite bias due to arbitrary choices of bounding parameters.

Regarding the second approach, the individual models may not be identifiable. Indeed, consider binary functional data where large regions of the domain are all 0 or 1. In these regions, a generalized additive model is identified only due to the smoothness assumption on the coefficient function $f_i(s)$, which may result in the divergence of the estimated log odds. Moreover, all smoothing assumptions are done at the individual level, without taking into account the information from the other subjects. In our data application, many participants are completely inactive during the very early morning hours (2AM-5AM). In contrast, the fast-GFPCA approach borrows strength from the other study participants to provide reasonable estimates at the subject level in areas where there is very little information for many, but not all, study participants.

In addition to the issues mentioned above, both approaches involve tuning parameters

related to the estimation method for individual model fits. For example, in the context of penalized splines, one must choose the number of splines, the basis, and a method for smoothing parameter selection. The latter point seems especially important given that the generative model implies the amount of “wiggliness” of each $i = 1, \dots, N$ latent function is the same. Fitting separate models does not enforce this constraint and may lead to some functions being estimated to be perfectly smooth, while others are quite wiggly. Moreover, while fast-GFPCA can deal with substantial missing areas in a data set, the two alternative methods described here cannot, especially when combined with excess zeros or ones.

However, the most important drawback of both these methods was that we were not able to successfully use them in our NHANES application; see an extensive discussion of these points in Section S-2 of the supplemental material. Moreover, fast-GFPCA can easily be extended to covariate adjusted and multilevel/structured generalized functional data. It is not immediately clear that either of the two approaches referenced above are readily extended to these scenarios.

3 Application

3.1 NHANES Accelerometry Data

The National Health and Nutrition Examination Survey (NHANES) is a large, ongoing study which provides a nationally representative sample of the non-institutionalized US population. NHANES is conducted by the Centers for Disease Control (CDC) and collects data in two-year waves with the goal of providing information on the health and nutrition of the US population. Wearable accelerometers were deployed in the 2003-2004, 2005-2006, 2011-2012, and 2013-2014 waves of NHANES.

The 2003-2006 accelerometry component of the NHANES study involved participants wearing a waist-worn accelerometer during waking hours. A guide to analyzing these data

is provided in Leroux et al. (2019) and an R data package, `rnhanesdata` Leroux (2022) is publicly available on Github at <https://github.com/andrew-leroux/rnhanesdata>. The 2011-2014 accelerometer data, released in 2021, are provided at multiple resolutions: subject, minute, and sub-second level. The subject and minute level data summarize individuals' acceleration patterns based on the new Monitor Independent Movement Summary (MIMS) unit John et al. (2019). Here, we use the 2011-2014 minute level MIMS data to construct active/inactive profiles for participants.

To obtain binary active/inactive profiles, we first threshold participants' daily MIMS data as $Y_{ih}^B(s) = 1\{Y_{ih}(s) \geq 10.558\}$, where $Y_{ih}(s)$ corresponds to the i^{th} individual's MIMS unit on day h at minute s . We then define their active/inactive profile as $Z_i(s) = \text{median}\{Y_{ih}^B(s) : h = 1, \dots, H_i\}$. For example, if $H_i = 7$, $Z_i(s)$ is 0 if study participant i was inactive at time s for at least 4 days and 1 otherwise. When the number of good days is even, say $\frac{H_i}{2}$, the median is defined as the $\frac{H_i}{2} + 1^{\text{th}}$ largest observation. The threshold for active/inactive on the MIMS unit scale is chosen to be 10.558, as suggested in Karas et al. (2022), though our methodology would apply similarly to other thresholds. The analytic sample contains data from $N = 4286$ participants with 1440 observations per person (minutes in a day).

Although NHANES is a nationally representative sample, obtaining nationally representative estimates for population quantities and model parameters requires the use of survey weights and survey design Korn and Graubard (2011); Lumley (2004); Skinner et al. (2017). The intersection of survey statistics and functional data analysis is a relatively new area of research Cardot et al. (2013a,b, 2014); Parker and Holan (2022) and is beyond the scope of the current work. Thus we do not account for the NHANES survey design in our data application, though it is an important direction for future methodological development.

3.2 Comparison methods and criteria

We apply fast-GFPCA (labeled `fastGFPCA`) to the NHANES data using bin widths (w_l) of 6, 10, and 30 minutes, and both overlapping and non-overlapping intervals. We compared methods to the fast binary variational FPCA (labeled `vbFPCA`) implementation in the `registr::bfpc()` function Wrobel (2018); Wrobel et al. (2019). We also consider a modified version of fast-GFPCA (labeled `modified fastGFPCA`) which further speeds up `fastGFPCA` in Step 4 by fitting the model (2) to four sub-samples of the data. The `modified fastGFPCA` approach is described in more detail in Section S-3 of the supplemental material. The approach described in Gertheiss et al. (2017) was not computationally feasible for the NHANES data. To facilitate comparisons across models, we fix the number of principal components across all methods to $K = 4$.

For each approach we compare model parameters and predictive performance. For model parameters we compare the first four estimated eigenfunctions and the population mean function. For predictive performance, we compare the estimated in-sample log-loss associated with each model fit and the estimated area under the receiver operating curve (AUC). Finally, we compare computation times across methods. Though Step 2 of fast-GFPCA could easily be parallelized, computation times are reported for serialized fitting of the models to provide an upper bound for this step.

Substantial differences were identified in terms of estimated population means and eigenfunctions using the `vbFPCA` approach and the fast-GFPCA approach (see commentary in Section 3.3). To investigate these differences a brief simulation study based on the NHANES data was conducted, described and summarized in Section S-4.1 of the supplemental material.

3.3 Results

3.3.1 Data Application

Model Parameters. Figure 2 displays the estimated population means $\hat{\beta}_0(s)$ (Figure 2A), and first four eigenfunctions $\hat{\phi}(s)$ (Figure 2B) of the latent process. Each column of Figure 2 corresponds to a different model fit where color indicates approach (vbFPCA in red, columns 1-2, and fast-GFPCA in blue, columns 3-8). Modified fast-GFPCA estimates a population mean function for each of four randomly selected sub-samples, which are then averaged to produce the overall $\hat{\beta}_0(s)$. These sub-sample estimates of $\beta_0(s)$ are shown as dashed lines in Figure 2A, columns 3-8. As the modified fast-GFPCA uses the population level $\hat{\phi}(s)$, there are no corresponding dashed plots for the eigenfunctions.

The population mean is fairly stable across sub-samples for the modified fast-GFPCA approach (Figure 2A, columns 3-8). We also find excellent agreement between the estimated linear predictor of the modified and unmodified fast-GFPCA approaches; see Figure 3 and associated discussion. Moreover, $\hat{\beta}_0(s)$ and $\hat{\phi}(s)$ are similar across the fast-GFPCA fits (overlapping vs non-overlapping windows, and bin widths), suggesting that the fast-GFPCA algorithm is fairly robust to the choice of input parameters in the NHANES data.

However, there are substantial differences between vbFPCA and fast-GFPCA both in terms of the estimated population mean functions and first three eigenfunctions. The largest differences in $\hat{\beta}_0(s)$ occur around 6AM-8AM and 9PM-12AM, where vbFPCA respectively over- and under-estimates $\beta_0(s)$ relative to fast-GFPCA. The vbFPCA $\hat{\beta}_0(s)$ suggests that, for participants with $b_i(s) = 0$, the probability of being active between 6AM-8AM is around 80-90% compared to around 50% from fast-GFPCA. This result is unexpected and does not match the results in the data, where we observe approximately 50% probability of being active during this time. One potential explanation could be that vbFPCA provides biased estimators of the mean and that the bias may be shifted into the latent random process variation. This hypothesis appears to be supported by our data

driven simulation in Section S-4.1 of the supplemental material. It is unclear at this time what is driving the observed bias in the vbFPCA approach as the same behavior is not seen in simulations presented in Section 4, but it may be related to the high proportion of observed 0s (inactive periods) during the night time hours in our data application.

Linear Predictor. Our goal in this section is to understand how differences in estimated model parameters $\hat{\beta}_0(s)$ and $\hat{\phi}(s)$ across methods lead to differences in the estimated linear predictor $\hat{\eta}_i(s); i \in 1, \dots, N$. To capture this, for each minute of the day we regressed $\hat{\eta}_i(s)$ from fast-GFPCA with non-overlapping windows and $w = 6$ on $\hat{\eta}_i(s)$ from each of four other GFPCA models. Figure 3 shows results of these time-specific linear regressions. These have the form $E[Y] = \mathcal{B}_0 + \mathcal{B}_1 X$, where Y is $\hat{\eta}_i(s)$ from fast-GFPCA with no overlap windows and $w = 6$, and X is $\hat{\eta}_i(s)$ estimated by: (1) fast-GFPCA with non-overlapping windows and $w = 30$ (orange lines); (2) modified fast-GFPCA with non-overlapping windows and $w = 6$ (blue lines); (3) vbFPCA with $Kt=8$ (green lines); and (4) vbFPCA with $Kt=30$ (yellow lines). The first panel and second panels display regression coefficients $\hat{\mathcal{B}}_0$ and $\hat{\mathcal{B}}_1$, respectively, and the third panel displays the percent variance explained (R^2). The black dashed line corresponds to perfect agreement between Y and X .

We observe almost perfect agreement between the reference model (fast-GFPCA with no overlap and $w = 6$) and the corresponding modified fast-GFPCA model (blue lines, intercept ≈ 0 , slope ≈ 1 , left to right panels, respectively). Similarly, we observe near perfect linear association between the fast-GFPCA model with larger bin width $w = 30$ (orange lines) across the day ($R^2 \geq 0.95$, right panel), with a nearly 1-to-1 relationship (intercept ≈ 0 , slope ≈ 1) during the active hours of the day ($\approx 10AM$ to $8PM$). During the active hours of the day, the vbFPCA results suggest a similar 1-to-1 trend in the mean, though the lower R^2 suggests less agreement between the fast-GFPCA and vbFPCA approaches than across fast-GFPCA estimates with different input parameters.

Predictive Accuracy. The observed differences in the estimated linear predictor do not appear to translate into different predictive accuracy in terms of either AUC or log loss; see the two rightmost columns of Table 1. A possible explanation of the similar predictive accuracy but different estimation performance may be that the disagreement in predictions between models occurs primarily during the nighttime hours (Figure 3), when the probability of an individual being active is generally very low.

Computation Time. Computation times for each model fit are presented in the middle of Table 1 for fast-GFPCA (top rows) and the vbFPCA approach (bottom rows). For fast-GFPCA, computation times are broken down by step of the algorithm. The fast-GFPCA algorithm (top rows, “Modified” = “No”) requires about 3-4 hours while vbFPCA requires 7 minutes for $K_t=8$ and 20 minutes for $K_t=30$. These computation times are driven by Step 4, which is unavoidable if the model is fit on the entire NHANES data set. Indeed, the fast-GFPCA approach here is the simplest GLMM that can be fit while maintaining the functional structure of the data. When Step 4 is modified (top rows, “modified” = “Yes”) computation times decrease substantially and becomes comparable to the vbFPCA approach. From a practical perspective we have not found any substantial differences between the results obtained from the fast-GFPCA and modified fast-GFPCA approaches.

4 Simulation Study

Our simulations are designed to (1) quantify the computational efficiency and scalability of fast-GFPCA as sample and grid size increase, (2) evaluate the accuracy of our method in comparison with existing approaches for generalized and binary FPCA, and (3) understand the behavior of our method under different data binning strategies from Step 1 of our estimation algorithm in Section 2.2. For larger simulated datasets we focus on the binary functional data setting because the only existing competing approach that is computationally

ally feasible for large functional datasets is tailored specifically to binary data. For smaller datasets we also evaluate our method in comparison to an existing approach for GFPCA of Poisson functional data.

4.1 Simulation Design

We simulate binary functional data for $N = 100, 500, 1000$ subjects observed on a length $J = 100, 500, 2000$ grid in $S = [0, 1]$ that is equally spaced and shared across subjects. Poisson functional data are simulated from $N = 50, 100$ and $J = 100, 200$ due to the computational limitations of competing methods. Curves for $i \in 1, \dots, N$ subjects are drawn from model (2). We construct $K = 4$ principal components, with true eigenvalues $\lambda_k = 0.5^{k-1}; k = 1, 2, 3, 4$. The true eigenfunctions are drawn from one of two scenarios intended to mimic real data settings. In the first setting, latent curves $\eta_i(s)$ and eigenfunctions are periodic, with $\phi(s) = \{\sqrt{2} \sin(2\pi s), \sqrt{2} \cos(2\pi s), \sqrt{2} \sin(4\pi s), \sqrt{2} \cos(4\pi s)\}$. The second setting does not exhibit periodicity, with true eigenfunctions given by $\phi(s) = \{1, \sqrt{3}(2s - 1), \sqrt{5}(6s^2 - 6s + 1), \sqrt{7}(20s^3 - 30s^2 + 12s - 1)\}$. For most simulation settings we assume $\beta_0(s) = 0$, however, for a subset of scenarios we construct nonzero $\beta_0(s)$ using a B-spline basis, specifically for binary functional data with: (1) $N, J = (1000, 2000)$; and (2) $N, J = (500, 100)$.

For binary and Poisson data, $g(\cdot)$ is taken to be the logit and log links, respectively. For each subject and time point, exponential family observations $Z_i(s)$ are sampled independently from either a Bernoulli distribution with probability $\text{logit}^{-1}[\eta_i(s)]$, or from a Poisson distribution with rate $e^{[\eta_i(s)]}$.

4.2 Comparison to Existing Approaches

4.2.1 Binary Functional Data

We assess the performance of fast-GFPCA across different bin widths w_l and compare non-overlapping and overlapping windows. Specifically, for each simulation scenario we evaluate three different bin widths, $w_l \in (6, 10, 50)$. We do not estimate fast-GFPCA with $w_l = 50$ when $J < 500$ as the large bin size does not make sense in this context.

We compare fast-GFPCA with two different binary FPCA approaches, both of which are implemented in the `registr` package Wrobel (2018); Wrobel and Bauer (2021). The first method to which we compare is the two-step conditional GFPCA model introduced by Gertheiss et al. (2017), which is implemented using the `registr::gfPCA_twoStep()` function and referred to as `tsGFPCA` in text and figures below. While `gfPCA_twoStep()` is a general purpose function that can accommodate multiple exponential family distributions, it is computationally intensive and thus impractical for most simulation settings. To reduce this computational burden we only implement `tsGFPCA` when $N \in (100, 500)$ and $J = 100$. We compare fast-GFPCA in all binary data simulation settings to the `vbFPCA` algorithm from Wrobel et al. (2019), which is implemented via the `registr::bfPCA()` function, and is denoted `vbFPCA`. This method is highly computationally efficient, but designed for binary functional data with a logit link, and cannot be generalized to other link functions or exponential family distributions. Because the `vbFPCA` approach models population mean $\beta_0(s)$ and eigenfunctions $\phi(s)$ using a B-spline expansion without a smoothness penalty, the number of basis functions must be manually tuned to obtain optimal smoothness. To address this, for each simulated dataset we implement `vbFPCA` with $Kt = 8$ basis functions (the package default) and $Kt = 30$.

For both competing methods we implement a periodic B-spline basis using the `registr` option `periodic = TRUE`. By default in the `registr` package, eigenfunctions $\phi(s)$ are returned unscaled and on a grid of size 100. To enable comparison with results from fast-

GFPCA, we linearly interpolate eigenfunctions estimated using `tsGPCA` and `vbFPCA` to a grid of size J and scale by the square root of the grid length.

4.2.2 Poisson Functional Data

In the Poisson setting we compare fast-GFPCA with `registr::gfpca_twoStep()`. Since the comparative method is highly computationally intensive, we only consider small data settings of sample sizes $N \in (50, 100)$, grid lengths $J \in (100, 200)$. We simulate 100 datasets for each of the six simulation scenarios arising from this combination of grid length and sample size. For fast-GFPCA we compare non-overlapping and overlapping windows, and consider bin widths $w_l \in (6, 10)$.

4.3 Evaluation Criteria

We compare the performance of the three methods (`tsGFPCA`, `vbFPCA`, and `fastGFPCA`) with respect to accuracy in recovering subject-specific latent means in the linear predictor space $\eta_i(s)$, accuracy in recovering population-level mean $\beta_0(s)$ and eigenfunctions $\phi(s)$, and computational efficiency. Accuracy of subject-specific log-odds across models is quantified using mean integrated squared error (MISE) across subjects, given by $\frac{1}{N} \sum_{i=1}^N \int_0^1 (\hat{\eta}_i(s) - \eta_i(s))^2 ds$. Accuracy of eigenfunction estimation is compared using MISE defined by $\frac{1}{k} \sum_{k=1}^4 \int_0^1 (\hat{\phi}_k(s) - \phi_k(s))^2 ds$, and population mean accuracy is measured using ISE. Computation times are reported in minutes.

4.4 Simulation Results: Accuracy

Tables 2, 3, and 4 summarize accuracy of key quantities across methods and simulation scenarios, where outcome data were generated from both binomial and Poisson distributions with periodic true eigenfunctions. Table 2 provides the MISE for $\hat{\eta}_i(s)$, the estimated subject-specific latent means in the linear predictor space, which are log-odds for binomial

data and log-rates for Poisson data. Similarly, Tables 3 and 4 summarize the MISE of eigenfunctions $\phi(s)$ and ISE of population-level mean $\beta_0(s)$, respectively.

4.4.1 Binomial data

Table 2 shows that all methods estimate the latent log-odds accurately. Our **fastGFPCA** approach performs equally well to the best competing method, **vbFPCA** with $Kt = 8$ spline bases in all but one scenario. Simulated data are periodic, and as a result **fastGFPCA** with overlapping bins outperforms **fastGFPCA** with non-overlapping bins. **fastGFPCA** performs best for the largest bin width, $w_l = 50$, except when the grid size J is smallest ($J = 100$). However, **fastGFPCA** performs well across all chosen bin widths.

Table 3 indicates that our **fastGFPCA** approach recovers true population eigenfunctions comparably to or better than the competing **vbFPCA** method in every scenario. For smaller grid sizes J the **vbFPCA** approach with $Kt = 8$ performs slightly better, but **fastGFPCA** outperforms **vbFPCA** when J increases. Table 4 shows that **fastGFPCA** outperforms **vbFPCA** at recovering the population mean $\beta_0(s)$ in all simulation scenarios. This is likely due to the fact that **fastGFPCA** penalizes spline coefficients to obtain a flat line estimate around the true value of $\beta_0(s) = 0$ while **vbFPCA**, which does not penalize smoothness, cannot estimate $\beta_0(s)$ as well when the true function is linear. Similar results for non-periodic data are observed in Supplemental Table S1.

Figure 4 highlights these results at a more granular level for one simulation scenario with $N = 500$ subjects and $J = 500$ time points and a nonzero, nonlinear population mean function $\beta_0(t)$. Specifically, Figure 4 shows the estimated population mean function, $\hat{\beta}_0(s)$, and the first four estimated eigenfunctions, $\hat{\phi}_k(s); k \in 1, \dots, 4$, from 100 simulated datasets. Estimates are presented as red lines for models using the competing **vbFPCA** approach or blue lines for models using our proposed **fastGFPCA** method, with dotted black lines representing the true value. All methods provide reasonable results in this simulation setting. Our **fastGFPCA** approach with overlapping bins and bin width $w_l = 50$ provides

the best results. The `vbFPCA` method with $Kt = 8$ spline basis functions also performs well, though overestimates $\beta_0(t)$ at the beginning of the functional domain s . The `fastGFPCA` approach with non-overlapping bins and $w_l = 50$ overestimates $\phi_4(s)$ at the endpoints of the functional domain, which suggests that a smaller band width may be more appropriate when data is not periodic and only non-overlapping bins can be used for `fastGFPCA`.

4.4.2 Poisson data

For Poisson distributed functional data the proposed `fastGFPCA` method was compared with the two-step approach `tsGFPCA` because there is no variational EM method for Poisson FPCA. Table 3 indicates that `fastGFPCA` estimates the latent log-rate far better than the competing `tsGFPCA` method in every simulation scenario. Similarly, Tables 3 and 4 indicate that `fastGFPCA` recovers the true population mean and eigenfunctions much more accurately than `tsGFPCA` in every scenario. Of the `fastGFPCA` approaches, `fastGFPCA` with overlapping bins and $w_l = 10$ tends to perform best, but the different `fastGFPCA` models perform similarly regardless of window overlap or choice of bin width w_l .

Figure S2 in the supplemental material provides some intuition as to why `tsGFPCA` performs poorly in the Poisson setting. This figure shows the estimated population mean function and eigenfunctions from 25 simulated datasets with $N = 100$ subjects and $J = 100$ time points. Model estimates are presented as red lines (`tsGFPCA`) or blue lines (`fastGFPCA`). The proposed `fastGFPCA` method provides reasonable results for bin widths $w_l = 6$ and $w_l = 10$, but `tsGFPCA` clearly provides incorrect estimates.

4.5 Simulation Results: Computational Efficiency

Table 5 shows median computation time in minutes across methods and simulation scenarios for both binomial and Poisson functional data. Across all scenarios the non-overlapping `fastGFPCA` approach is more computationally efficient than the overlapping `fastGFPCA` approach. Bin widths w_l have a negligible effect on computation time for non-overlapping

bins, but when bins are overlapping computation time increases with increasing bin width. `fastGFPCA` scales well when grid size J increases but more slowly when one increases the number of subjects, N . Notably, at smaller sample sizes ($N \in \{100, 500\}$), `fastGFPCA` is comparably efficient or faster than `vbFPCA`, a method custom-built for speed. For $N \geq 1000$, `fastGFPCA` can be sped-up using techniques discussed in Supplemental Section S-3. The `tsGFPCA` has a median time of 92 minutes for Poisson data with just 100 subjects and 200 time points. This indicates that `tsGFPCA` is prohibitively slow for our data application of 4286 subjects with 1440 time points each.

5 Discussion

The fast-GFPCA method proposed in this manuscript represents a simple, understandable, and computationally feasible solution to the complex task of estimating functional principal components analysis for non-Gaussian data. In addition, we have provided a mathematical justification for the principles that underlie fast-GFPCA and shown that the method compares favorably to the few existing approaches in terms of both estimation accuracy and computational efficiency. Moreover, existing methods, such as the `vbFPCA` approach for binary data presented here may provide biased estimates of model parameters (population mean function and covariance operator), suggesting that the fast-GFPCA approach is a reasonable method for comparison even when an alternative approach is more computationally efficient in a given application.

Though the work here shows fast-GFPCA to be fast, accurate, and appropriate for analyzing the motivating NHANES data, methodologic work remains. Specifically, it is unclear at this time how to choose the optimal bin, both with regard to bin width and the decision to use overlapping versus non-overlapping windows for estimation. While a cross-validated prediction error criteria may be a viable option, subject-level cross-validation requires prediction of random effects in non-Gaussian models using participants' data not

included in model fitting, a non-trivial problem in non-Bayesian contexts. Moreover, when using non-overlapping windows, automated smoothing parameter selection of FPCA on the latent process in Step 3 of the fast-GFPCA algorithm is unreliable. Here we propose an ad-hoc solution based on visual inspection of the eigenfunctions and/or estimated covariance function. While this is feasible due to the speed of the FACE method implemented in Step 3, we would prefer a fully automated approach. Deriving an appropriate variation of the GCV criteria used by FACE for smoothing parameter selection which accounts for autocorrelated data may improve the method proposed here. Nevertheless, the results of this work represent an encouraging step forward for estimation of GFPCA in very high dimensional data, specifically large N , a key bottleneck for the application of functional data analysis methods in practice.

An appealing feature of fast-GFPCA is that it can be extended to: (1) covariate dependent GFPCA; (2) multilevel, structured, and longitudinal GFPCA.

Covariate dependent GFPCA. The fast-GFPCA method easily incorporates covariates into the model. Consider the case of one additional scalar predictor (e.g., age), denoted x_i . Step 2 of fast-GFPCA is simply modified to fit local models of the form $g(E[Z_i(s_j)|s_j \in S_l]) = \beta_0(s_{m_l}) + \beta_1(s_{m_l})x_i + b_i(s_{m_l}) = \eta_i(s_{m_l})$. Then, in Step 4, the final GLMM includes the additive term associated with the proposed varying coefficient model. The effort required for this extension is minimal.

Nested, longitudinal or crossed design GFPCA. Consider the case when multiple functions are observed per study participant. For example, in the NHANES data each study participant has multiple days of accelerometry data. For notation simplicity assume that there are K functions for every study participant. The extension to multilevel GFPCA follows naturally from the fast-GFPCA algorithm. Specifically, in Step 2 we fit the multilevel model $g(E[Z_{ik}(s_j)]) = \beta_0(s_{m_l}) + b_i(s_{m_l}) + v_{ik}(s_{m_l})$. Step 3 can then proceed with

MFPCA FACE Cui et al. (2022) to estimate the principal directions of variation at each level. If the functional data has longitudinal Greven et al. (2010) or crossed Shou et al. (2015) designs the local GLMM can be changed accordingly.

References

- Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, 54(4):280–288.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.
- Cardot, H., Dessertaine, A., Goga, C., Josserand, É., and Lardin, P. (2013a). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption. *Survey Methodology*, 39(2):283–301.
- Cardot, H., Goga, C., and Lardin, P. (2013b). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic journal of statistics*, 7:562–596.
- Cardot, H., Goga, C., and Lardin, P. (2014). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. *Scandinavian Journal of Statistics*, 41(2):516–534.
- Chen, H., Wang, Y., Paik, M. C., and Choi, H. A. (2013). A marginal approach to reduced-rank penalized spline smoothing with application to multilevel functional data. *Journal of the American Statistical Association*, 108(504):1216–1229.
- Chiou, J.-M., Chen, Y.-T., and Yang, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, 24(4):1571–1596.
- Cui, E., Li, R., Crainiceanu, C. M., and Xiao, L. (2022). Fast multilevel functional principal component analysis. *Journal of Computational and Graphical Statistics*, 0(ja):1–33.
- Di, C., Crainiceanu, C., Caffo, B., and Punjabi, N. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics*, 3(1):458–488.
- Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., van Hees, V. T., Trenell, M. I., Owen, C. G., Preece, S. J., Gillions, R., Sheard, S., Peakman, T., Brage, S., and Wareham, N. J. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: The uk biobank study. *PLOS ONE*, 12(2):1–14.

- Gertheiss, J., Goldsmith, J., and Staicu, A.-M. (2017). A note on modeling sparse exponential-family functional response curves. *Computational Statistics & Data Analysis*, 105:46–52.
- Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. (2020). *refund: Regression with Functional Data*. R package version 0.1-23.
- Goldsmith, J., Zipunnikov, V., and Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71(2):344–353.
- Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, pages 1022–1054.
- Hall, P., Müller, H.-G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):703–723.
- Happ, C. and Greven, S. (2015). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113:649 – 659.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520.
- J., P. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, 18(4):995–1015.
- James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.
- John, D., Tang, Q., Albinali, F., and Intille, S. (2019). An open-source monitor-independent movement summary for accelerometer data processing. *Journal for the Measurement of Physical Behaviour*, 2(4):268 – 281.
- Jolliffe, I. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society, Series C*, 31(3):300–303.
- Jones, M. C. and Rice, J. A. (1992). Displaying the important features of large collections of similar curves. *The American Statistician*, 46(2):140–145.
- Karas, M., Muschelli, J., Leroux, A., Urbanek, J. K., Wanigatunga, A. A., Bai, J., Crainiceanu, C. M., and Schrack, J. A. (2022). Comparison of accelerometry-based measures of physical activity: Retrospective observational data analysis study. *JMIR Mhealth Uhealth*, 10(7):e38077.
- Korn, E. L. and Graubard, B. I. (2011). *Analysis of health surveys*, volume 323. John Wiley & Sons.

Leroux, A. (2022). *rnhanesdata: NHANES Accelerometry Data Pipeline*. R package version 1.02.

Leroux, A., Di, J., Smirnova, E., McGuffey, E. J., Cao, Q., Bayatmokhtari, E., Tabacu, L., Zipunnikov, V., Urbanek, J. K., and Crainiceanu, C. (2019). Organizing and analyzing the activity data in nhanes. *Statistics in Biosciences*, 11(2):262–287.

Li, C. and Xiao, L. (2021). *mfaces: Fast Covariance Estimation for Multivariate Sparse Functional Data*. R package version 0.1-3.

Li, Z. and Wood, S. N. (2020). Faster model matrix crossproducts for large generalized linear models with discretized covariates. *Statistics and Computing*, 30(1):19–25.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19.

Parker, P. A. and Holan, S. H. (2022). A bayesian functional data model for surveys collected under informative sampling with application to mortality estimation using nhanes. *Biometrics*, n/a(n/a).

Pearson, K. (1901). LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer New York, NY, USA.

Rice, J. and Silverman, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1):233–243.

Shou, H., Zipunnikov, V., Crainiceanu, C., and Greven, S. (2015). Structured functional principal component analysis. *Biometrics*, 71(1):247–257.

Skinner, C., Wakefield, J., et al. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, 32(2):165–175.

Staniswalis, J. and Lee, J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418.

van der Linde, A. (2008). Variational Bayesian functional PCA. *Computational Statistics & Data Analysis*, 53(2):517–533.

van der Linde, A. (2009). A Bayesian latent variable approach to functional principal components analysis with binary and count data. *AStA Advances in Statistical Analysis*, 93(3):307–333.

- Weishampel, A., Staicu, A.-M., and Rand, W. (2023). Classification of social media users with generalized functional data analysis. *Computational Statistics & Data Analysis*, 179:107647.
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.
- Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017). Generalized additive models for gigadata: Modeling the u.k. black smoke network daily data. *Journal of the American Statistical Association*, 112(519):1199–1210.
- Wood, S. N., Scheipl, F., and Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, 23(3):341–360.
- Wrobel, J. (2018). register: Registration for exponential family functional data. *Journal of Open Source Software*, 3(22):557.
- Wrobel, J. and Bauer, A. (2021). registr 2.0: Incomplete curve registration for exponential family functional data. *Journal of Open Source Software*, 6(61):2964.
- Wrobel, J., Park, S. Y., Staicu, A. M., and Goldsmith, J. (2016). Interactive graphics for functional data analyses. *Stat*, 5(1):108–118.
- Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. (2019). Registration for exponential family functional data. *Biometrics*, 75(1):48–57.
- Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2018). Fast covariance estimation for sparse functional data. *Statistics and Computing*, 28(3):511–522.
- Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2021). *face: Fast Covariance Estimation for Sparse Functional Data*. R package version 0.1-6.
- Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data. *Statistics and computing*, 26(1):409–421.
- Yao, F., Müller, H., Clifford, A., Dueker, S., Follett, J., Lin, Y., Buchholz, B., and Vogel, J. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685.
- Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 28(100):577–590.

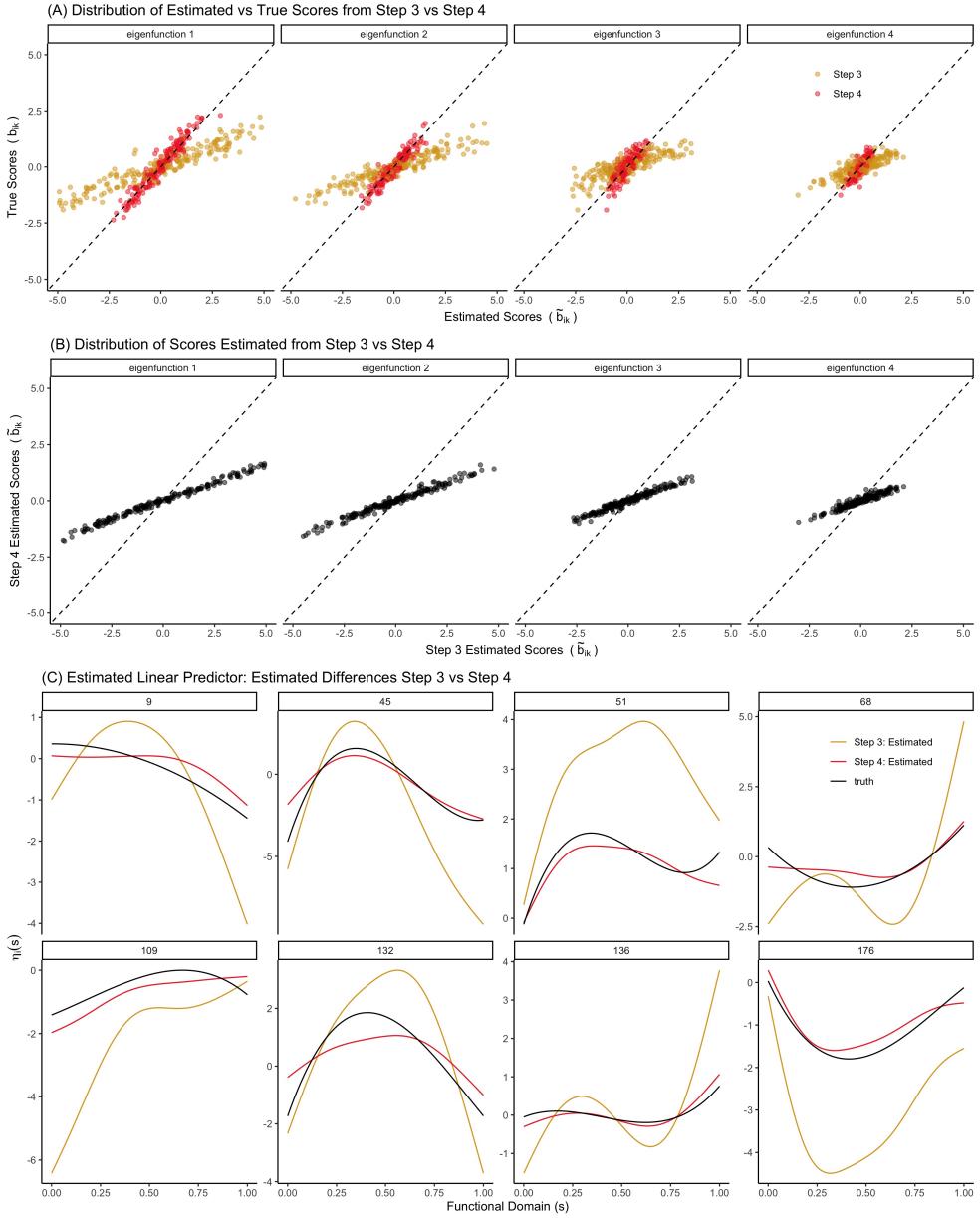


Figure 1: Illustration of the need for Step 4 in the fast-GFPCA method using one simulated dataset from the simulation study ($N = 200$, $J = 200$, $w_l = 10$, $\phi(s) = \{\sqrt{2} \sin(2\pi s), \sqrt{2} \cos(2\pi s), \sqrt{2} \sin(4\pi s), \sqrt{2} \cos(4\pi s)\}$). (A) Plot of the estimated scores from step 3 (golden points) and step 4 (red points) on the x-axis versus the true scores on the y-axis separately for each of the first four eigenfunctions. The black line represents the identity line. (B) Plot of the scores estimated from step 3 (x-axis) versus those from step 4 (y-axis). (C) Estimated curves on the linear predictor scale

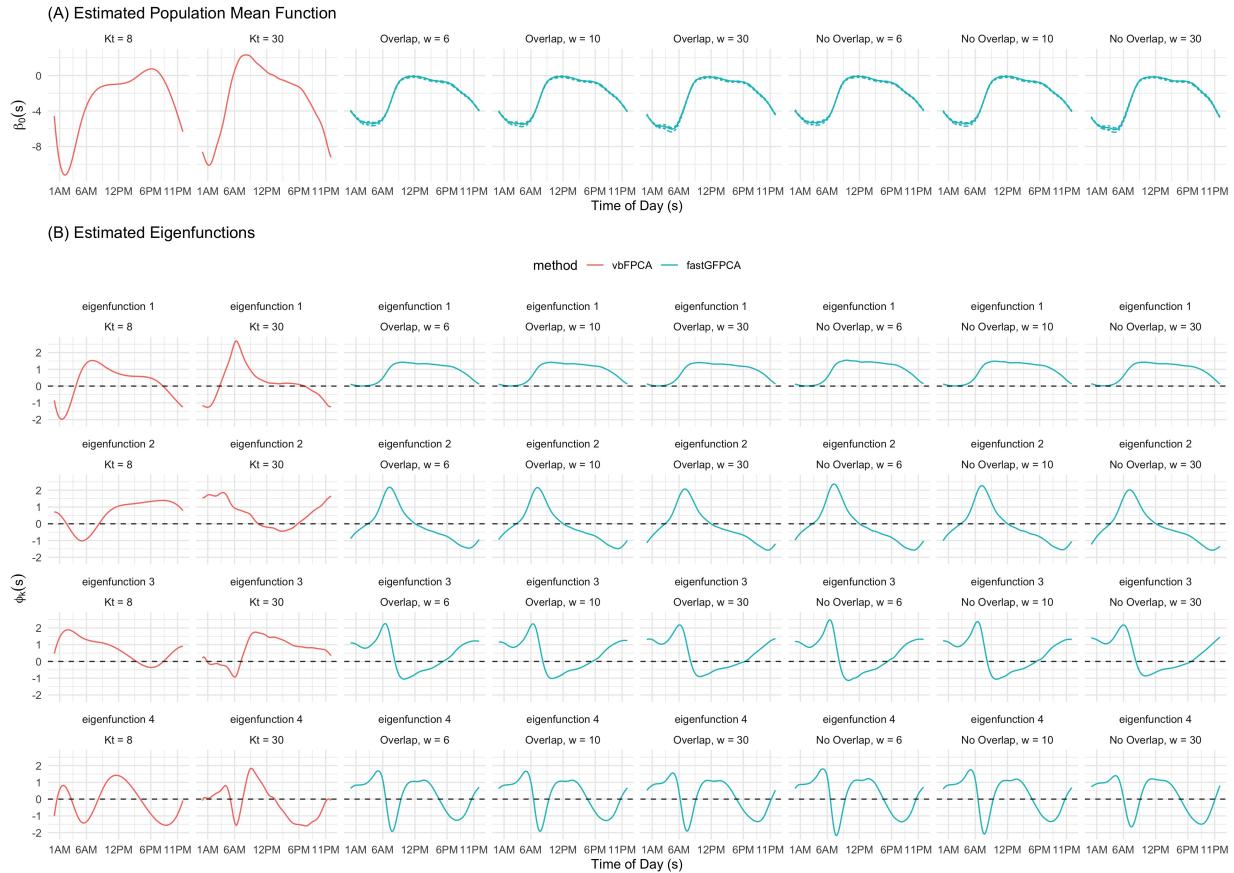


Figure 2: Estimated population mean function (first row) and the first four estimated eigenfunctions (rows 2 – 5) in NHANES. Model estimates are presented as red (vbFPCA) or blue lines (fast-GFPCA). The two leftmost columns correspond to vbFPCA ($K_t = 8$ in column 1 and $K_t = 30$ in column 2). The six rightmost columns correspond to fast-GFPCA with different input parameters (overlapping versus non-overlapping windows) and window sizes ($w = 6, 10, 30$). Estimates of the population mean function based on the modified fast-GFPCA are displayed as dashed lines.

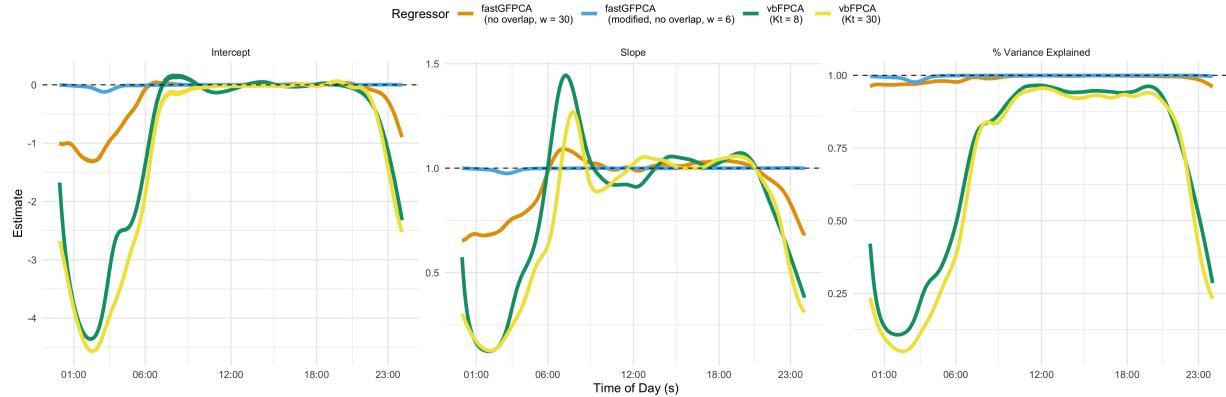


Figure 3: Results from time-specific regression of the estimated log odds of being active obtained from fast-GFPCA using $w_l = 6$ with non-overlapping windows and: (1) fast-GFPCA using $w_l = 30$ with non-overlapping windows (red lines); (2) vbFPCA with $K_t = 8$ (green lines); and (3) vbFPCA with $K_t = 30$ (blue lines). Regressions of the form $E[Y] = \mathcal{B}_0 + \mathcal{B}_1 X$ were fit separately for each minute, with the resulting estimates $\hat{\mathcal{B}}_0$, $\hat{\mathcal{B}}_1$, and the percent variance explained (R^2) plotted separately in each panel (left to right). Deviations from the black dashed line in each panel denote a reduced rate of agreement between the regressor and the results from fast-GFPCA using $w_l = 6$ with non-overlapping windows.

fastGFPCA								
Parameters	Modified	Computation Time (mins)					AUC	Log-loss
		Step 1	Step 2	Step 3	Step 4	Total		
Overlap, $w = 6$	No		16.42	0.04	174.20	190.63	0.909	0.328
	Yes		16.42	0.04	10.90	27.33	0.909	0.328
Overlap, $w = 10$	No	23.28	0.04	184.82	208.12	0.909	0.327	
	Yes	23.28	0.04	10.74	34.03	0.909	0.327	
Overlap, $w = 30$	No	55.90	0.04	184.66	240.57	0.909	0.327	
	Yes	55.90	0.04	10.48	66.38	0.909	0.327	
No Overlap, $w = 6$	No	2.48	0.01	188.90	191.43	0.909	0.328	
	Yes	2.48	0.01	11.34	13.13	0.909	0.328	
No Overlap, $w = 10$	No	2.19	0.01	185.04	187.28	0.909	0.327	
	Yes	2.19	0.01	10.89	13.13	0.909	0.327	
No Overlap, $w = 30$	No	1.97	0.01	197.35	199.36	0.909	0.327	
	Yes	1.97	0.01	11.60	13.61	0.909	0.327	

vbFPCA (registr::bfPCA())							
Parameters	Computation Time (mins)				AUC	Log-loss	
Kt=8					7.11	0.909	0.328
Kt=30					20.86	0.910	0.326

Table 1: Computation times and in-sample predictive performance summaries for **fastGFPCA** (top rows) estimated under different parameter settings (column 1, overlapping versus non-overlapping windows, window size of $w = 6, 10, 30$) using both the primary and modified algorithm (column 2, Modified = No or Yes, respectively), and variational Bayes (vbFPCA, bottom rows) estimated with either $K_t = 8$ or $K_t = 30$. For **fastGFPCA**, computation times are presented using both the total time and separately by each step of the procedure (step 1 is left blank as computationally this step is effectively instantaneous). AUC and log-loss, presented in the rightmost two columns are calculated over all minutes of the day.

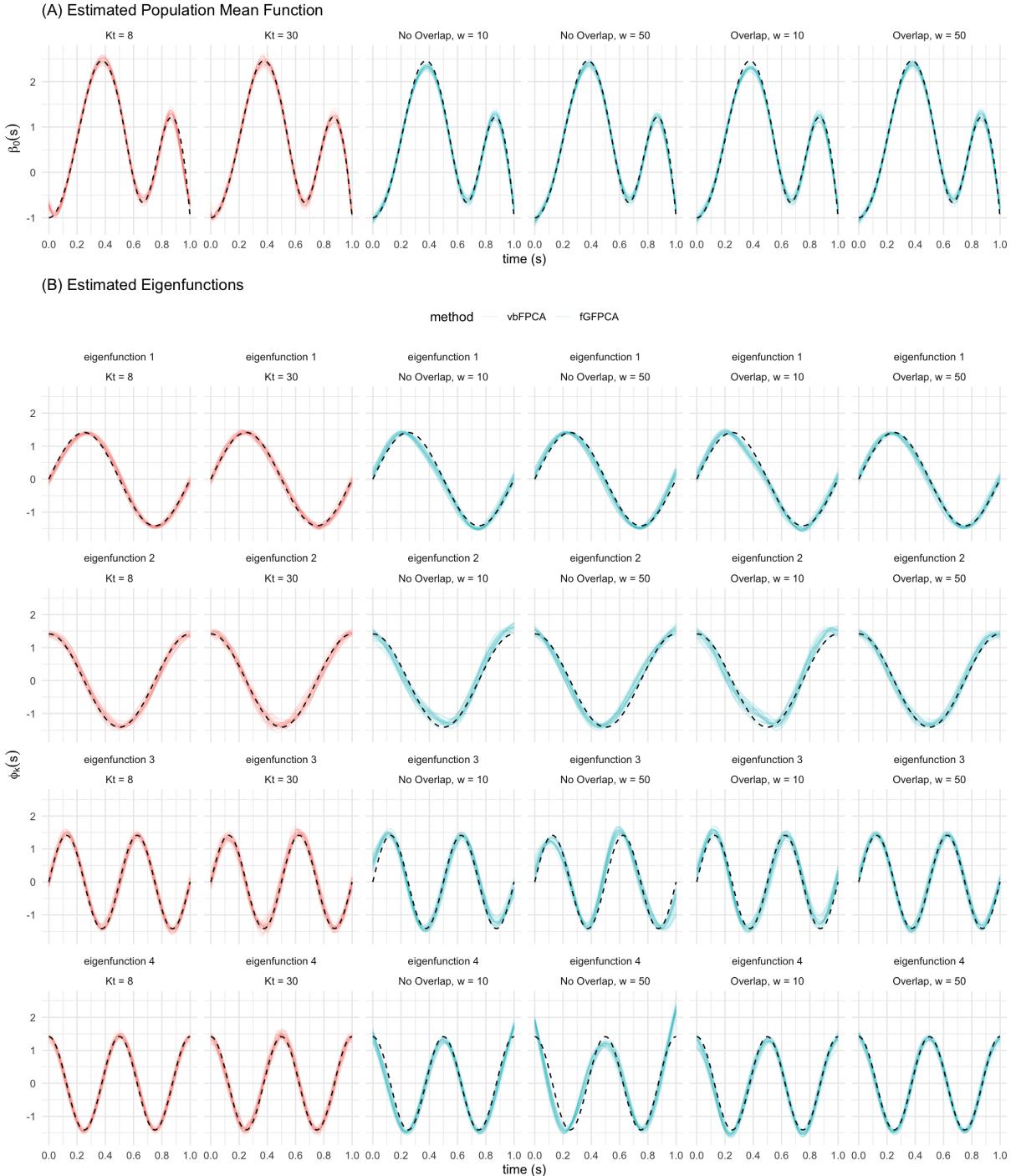


Figure 4: The estimated population mean function $\hat{\beta}_0(s)$ and eigenfunctions $\hat{\phi}(s)$ from 100 simulated binary functional datasets with $N = 500$ subjects and $J = 500$ time points. Model estimates are red lines (**vbFPCA**) or blue lines (**fastGPPCA**). **vbFPCA** estimates from models using either $Kt = 8$ (left column) or $Kt = 30$ (second column) basis functions are compared with six **fastGPPCA** models. From left to right, the **fastGPPCA** models (blue columns) are estimated with non-overlapping windows with bin sizes $w_l = 10$ and $w_l = 50$, and overlapping windows with bin sizes $w_l = 10$ and $w_l = 50$.

Mean Integrated Squared Error of $\eta_i(s)$											
Family	N	J	Fast-GFPCA						vbFPCA		tsGFPCA
			Overlapping Bins			Non-Overlapping Bins			K = 8	K = 30	K = 30
			$w_l = 6$	$w_l = 10$	$w_l = 50$	$w_l = 6$	$w_l = 10$	$w_l = 50$			
Binomial	100	100	2.19	2.14	-	2.39	2.44	-	2.15	2.87	2.18
		500	0.55	0.52	0.49	0.53	0.53	0.58	0.49	0.63	-
		2000	0.17	0.15	0.13	0.15	0.15	0.13	0.13	0.16	-
	500	100	2.05	2.02	-	2.33	2.43	-	2.02	2.15	2.03
		500	0.5	0.49	0.47	0.5	0.5	0.58	0.47	0.5	-
		2000	0.15	0.14	0.12	0.14	0.14	0.13	0.12	0.13	-
	1000	100	2.05	2.04	-	2.33	2.41	-	2.01	2.08	-
		500	0.5	0.48	0.47	0.51	0.5	0.58	0.47	0.48	-
		2000	0.14	0.13	0.12	0.14	0.13	0.13	0.12	0.12	-
Poisson	50	100	0.52	0.49	-	0.56	0.65	-	-	-	13.0
		200	0.32	0.28	-	0.32	0.29	-	-	-	15.0
	100	100	0.45	0.43	-	0.5	0.59	-	-	-	15.9
		200	0.25	0.24	-	0.27	0.25	-	-	-	20.5

Table 2: Mean integrated squared error (MISE) for $\hat{\eta}_i(s)$, the estimated subject-specific latent means in the linear predictor space across methods and simulation scenarios. In each row the method(s) with the lowest MISE for that simulation scenario is in **bold**. All data summarized in this table were simulated with population mean $\beta_0(s) = 0$. An “-” indicates that that model was not evaluated for a given simulation scenario. All values are multiplied by a factor of 10.

Mean Integrated Squared Error of Eigenfunctions $\phi(s)$											
Family	N	J	Fast-GFPCA						vbFPCA		tsGFPCA
			Overlapping Bins			Non-Overlapping Bins			K = 8	K = 30	K = 30
			$w_l = 6$	$w_l = 10$	$w_l = 50$	$w_l = 6$	$w_l = 10$	$w_l = 50$			
Binomial	100	100	0.91	0.66	-	1.13	1.56	-	0.47	1.13	0.48
		500	0.66	0.55	0.46	0.57	0.59	0.58	0.28	0.39	-
		2000	0.64	0.54	0.42	0.55	0.52	0.42	0.72	0.88	-
	500	100	0.19	0.18	-	0.58	0.8	-	0.12	0.26	0.13
		500	0.13	0.11	0.08	0.13	0.11	0.32	0.08	0.11	-
		2000	0.14	0.12	0.1	0.13	0.13	0.1	0.56	0.62	-
	1000	100	0.11	0.08	-	0.5	0.7	-	0.05	0.11	-
		500	0.06	0.04	0.03	0.07	0.07	0.26	0.05	0.07	-
		2000	0.05	0.04	0.03	0.05	0.04	0.04	0.52	0.66	-
Poisson	50	100	1.03	0.94	-	1.16	1.24	-	-	-	10.97
		200	0.96	0.88	-	0.91	0.98	-	-	-	8.8
	100	100	0.4	0.34	-	0.52	0.66	-	-	-	10.79
		200	0.37	0.34	-	0.4	0.41	-	-	-	8.7

Table 3: Mean integrated squared error (MISE) for estimated population-level latent eigenfunctions, $\phi_k(s); k = 1, \dots, 4$. In each row the method(s) with the lowest MISE for that simulation scenario is in **bold**. All data summarized in this table were simulated with population mean $\beta_0(s) = 0$. An “-” indicates that that model was not evaluated for a given simulation scenario. All values are multiplied by a factor of 10.

Integrated Squared Error of Population Mean $\beta_0(s)$											
Family	N	J	Fast-GFPCA						vbFPCA		tsGFPCA
			Overlapping Bins			Non-Overlapping Bins			K = 8	K = 30	K = 30
			$w_l = 6$	$w_l = 10$	$w_l = 50$	$w_l = 6$	$w_l = 10$	$w_l = 50$			
Binomial	100	100	1.34	1.35	-	2.42	1.5	-	15.69	27.37	1.85
		500	0.53	0.54	0.48	0.42	0.42	0.62	12.94	15.88	-
		2000	0.1	0.09	0.11	0.06	0.08	0.07	12.13	12.11	-
	500	100	0.19	0.17	-	0.3	0.41	-	3.53	5.74	0.46
		500	0.06	0.06	0.06	0.05	0.05	0.06	3.03	3.45	-
		2000	0.02	0.02	0.01	0.02	0.02	0.02	3.12	3.13	-
	1000	100	0.1	0.09	-	0.14	0.12	-	1.69	2.87	-
		500	0.04	0.04	0.04	0.03	0.05	0.04	1.57	1.77	-
		2000	0.02	0.02	0.02	0.02	0.02	0.01	2.06	1.87	-
Poisson	50	100	4.34	3.46	-	5.12	5.54	-	-	-	116.59
		200	2.43	2.18	-	3.68	3.45	-	-	-	126.52
	100	100	2.31	2.18	-	3.27	4.81	-	-	-	174.8
		200	1.64	1.67	-	4.1	2.36	-	-	-	148.47

Table 4: Integrated squared error (ISE) for estimated population-level latent mean, $\beta_0(s)$. In each row the method(s) with the lowest ISE for that simulation scenario is in **bold**. All data summarized in this table were simulated with population mean $\beta_0(s) = 0$. An “-” indicates that that model was not evaluated for a given simulation scenario. All values are multiplied by a factor of 10^3 .

Median Computation Times											
Family	N	J	Fast-GFPCA						vbFPCA		tsGFPCA
			Overlapping Bins			Non-Overlapping Bins			K = 8	K = 30	K = 30
			$w_l = 6$	$w_l = 10$	$w_l = 50$	$w_l = 6$	$w_l = 10$	$w_l = 50$			
Binomial	100	100	0.1	0.1	-	0.1	0	-	0	0.2	16.9
		500	0.5	0.5	1.6	0.1	0.1	0.1	0.1	0.7	-
		2000	2	2.2	6.5	0.3	0.2	0.2	0.9	4.7	-
	500	100	2.3	2.3	-	2.2	2.2	-	0.1	0.7	71.3
		500	2.7	3.2	7	2	1.9	1.9	0.6	2.8	-
		2000	7.4	9.5	27	3.1	3.1	3.1	5.7	24.4	-
	1000	100	20.4	20.1	-	20.2	20.4	-	0.4	1.3	-
		500	18.6	20	29.9	17.2	17.4	17.2	1.7	6.4	-
		2000	30.1	35.1	75.9	22.3	22.6	22.4	21	71.4	-
Poisson	50	100	0.1	0.1	-	0.02	0.02	-	-	-	11.2
		200	0.2	0.21	-	0.04	0.03	-	-	-	40.4
	100	100	0.15	0.16	-	0.06	0.06	-	-	-	34.6
		200	0.25	0.27	-	0.07	0.06	-	-	-	92.0

Table 5: Median computation time in minutes across methods and simulation scenarios. In each row the method(s) with the fastest computation time for that simulation scenario is in **bold**. All data summarized in this table were simulated with population mean $\beta_0(s) = 0$. An “-” indicates that that model was not evaluated for a given simulation scenario.

Supplemental Material

S-1 Algorithmic Presentation of fast-GFPCA

Algorithm 1: Fast-GFPCA

Step 1: Bin the Data

Input : $\{Z_i(s_j)\}, 1 \leq i \leq N, 1 \leq j \leq J$
Output : $\{Z_i(s_j), l\} 1 \leq i \leq N, j \in \mathcal{S}_l, 1 \leq l \leq L\}$

Step 2: Estimate Local GLMMs

Input : $\{Z_i(s_j), l\} 1 \leq i \leq N, j \in \mathcal{S}_l, 1 \leq l \leq L\}$

Output : $\{\hat{\eta}_i(s_{m_l}), l\}, 1 \leq i \leq N, 1 \leq l \leq L$

for $l = 1, \dots, L$ **do**

 Fit a GLMM of the form:

$$g(E[Z_i(s_j)|s_j \in S_l]) = \beta_0(s_{m_l}) + b_i(s_{m_l}) = \eta_i(s_{m_l})$$

$$\text{Obtain } \hat{\eta}_i(s_{m_l}) = \hat{\beta}_0(s_{m_l}) + \hat{b}_i(s_{m_l})$$

end

Step 3: FPCA on Local Latent Estimates

Input : $\{\hat{\eta}_i(s_{m_l}), l\}, 1 \leq i \leq N, 1 \leq l \leq L$

Output : $\{\hat{\phi}_k(s_{m_l}), l\}, 1 \leq l \leq L, 1 \leq k \leq K$

Estimate $\text{Cov}\{\hat{\eta}_i(u), \hat{\eta}_i(v)\}$ for $u, v = m_1, \dots, m_L$ using the FACE algorithm

Obtain $\hat{\phi}_k(s)$ for $s = m_1, \dots, m_L$

Step 4: Re-estimate Subject Specific Scores

Input : $\{\hat{\phi}_k(s_l), l\}, 1 \leq l \leq L, 1 \leq k \leq K, \{Z_i(s_j)\}, 1 \leq i \leq N, 1 \leq j \leq J$

Output : $\{\hat{b}_i(s_j), \hat{\beta}_0(s_j)\}, 1 \leq i \leq N, 1 \leq j \leq J$

for $k = 1, \dots, K$ **do**

 | Obtain $\hat{\phi}_k(s)$ on the original grid s_1, \dots, s_J

end

Estimate GFPCA using additive generalized linear mixed models of the form

$$g(E[Z_i(s)]|\{\hat{\phi}_k(s) : 1 \leq s \leq J, 1 \leq k \leq K\}) = \beta_0(s) + b_i(s) = \beta_0(s) + \sum_{k=1}^K \xi_{ik} \hat{\phi}_k(s)$$

$$\xi_{ik} \stackrel{\text{iid}}{\sim} N(0, 2\sigma_k^2), \quad \text{Cov}(\xi_{ik}, \xi_{il}) = 0 \quad \text{for } k \neq l$$

S-2 Comparison to Weishampel in NHANES

Figure S1 presents the results of the GFPCA approach of Weishampel et al. (2023) versus fast-GFPCA applied to the NHANES data. We implemented the Weishampel et al. (2023) approach by first estimating individual models using 30 penalized cubic regression splines with REML smoothing parameter selection. FPCA on the latent scale was implemented using `refund::f pca.face()`. There are substantial differences in the shapes of these eigenfunctions (Figure S1(A)), with the first eigenfunction of Weishampel et al. (2023) (Figure S1(A), left panel, blue line) being dominated by the behavior at the tails of the functions (early AM and late PM). This is possibly due to the fact that many individuals' have all 0's (inactive) during the early AM (and to a lesser extent the late PM). When fitting individual GAMs to these data to estimate participants' latent functions, this tends to result in estimates which diverge at the boundary. This effect can be seen in the estimated log odds of being active obtained from Weishampel et al. (2023) in Figure S1(B). The approach of Weishampel et al. (2023) is compared to the results from fast-GFPCA for 8 randomly selected participants (each panel), with color indicating estimation approach (blue for Weishampel Weishampel et al. (2023), red for fast-GFPCA). Each of the estimates from the approach of Weishampel et al. (2023) diverges at the boundaries, though to varying extents.

S-3 Further speeding up fast-GFPCA

The primary potential computational bottleneck for Fast-GFPCA is the need to re-estimate the subject-specific scores in Step 4 of the Fast GFPCA algorithm. We have identified three potential avenues for speeding up the current algorithm, the first of which we apply in our data application, but leave detailed investigation of each to future work. The first approach is to estimate Steps 1-3 on the entire population, obtaining population level eigenfunctions for estimating subject-specific random effects, but estimating Step 4 on randomly selected

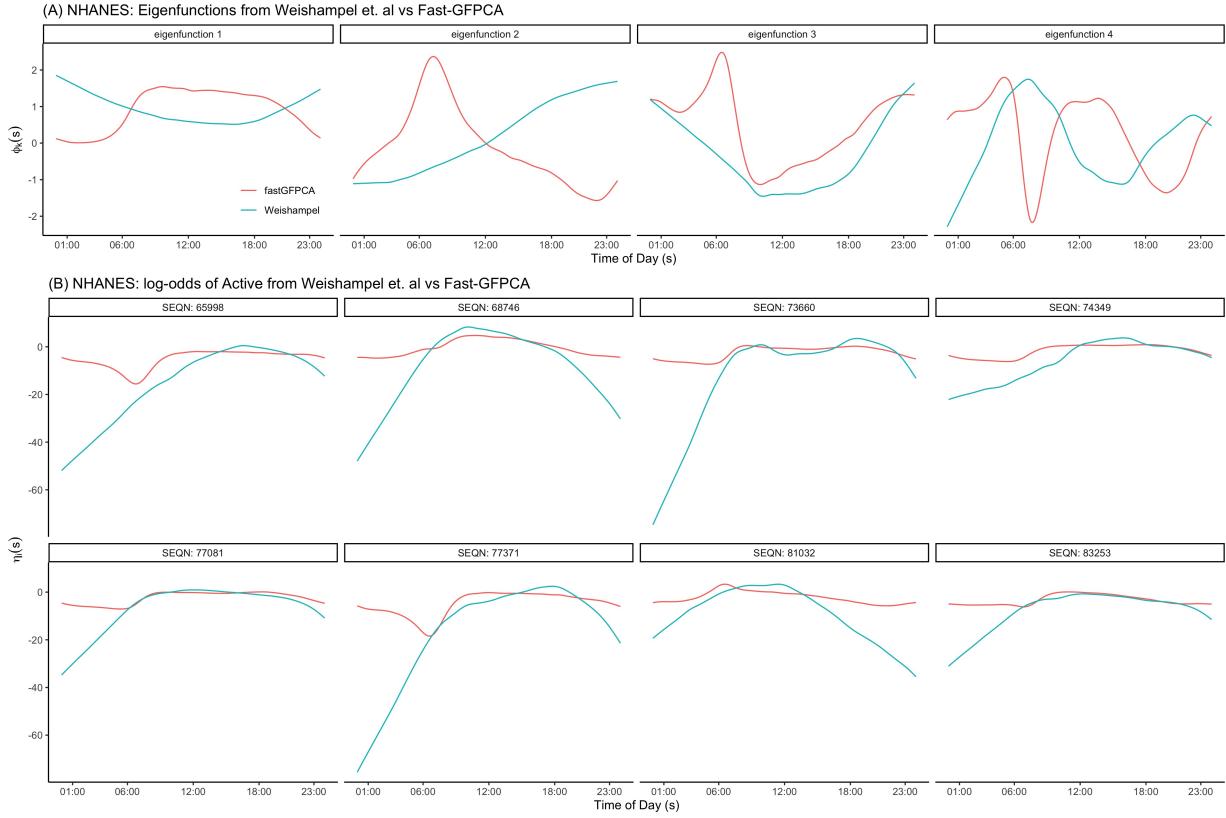


Figure S1: (A) The first four estimated eigenfunctions ($\hat{\phi}_k(s)$) from fast-GFPCA (red lines) versus those obtained from Weishampel et al (blue lines). (B) Estimated log-odds of being active Model estimates are presented as red lines (fast-GFPCA) or blue lines (fastGFPCA). **fastGFPCA** estimates compared with two **fastGFPCA** models. The **fastGFPCA** models (blue columns) are estimated with overlapping windows with bin sizes $w_l = 6$ and $w_l = 10$.

sub-samples of the data. In Step 4 all that is being done is re-estimating the population mean function and projection of the subject specific latent estimates onto the population basis estimates. For large sample sizes, such as in our data application, both of these quantities should be relatively stable in sufficiently large sub-samples (e.g. $N = 1000$).

The remaining two approaches are based on the observation in our simulations and data application that the scores obtained in Step 3 are almost perfectly correlated with those from Step 4, but are off by linear scaling factor (See Figure 1(B)). This scaling factor appears to be the same or similarly valued across each of the eigenfunctions Note the slope is nearly identical across eigenfunctions in Figure 1(B)). This suggests that if one were able to estimate the scaling factor, final estimates for the subject specific scores

would require less computation. We see this as achievable in at least two ways. First, one may estimate Step 4 on a subset of the data, then apply the scaling factor directly to the estimates obtained from Step 3. Alternatively, one may apply the previously suggested subset analysis to estimate the variance of subject scores, λ_k , and then treat both the estimated eigenfunctions ($\{\hat{\phi}_k(s)\}, 1 \leq s \leq J$) and the variance parameters as fixed. In this way, the model in Step 4 requires estimating only the population mean ($\beta_0(s)$), substantially reducing the computational complexity of Step 4. This can be done easily in the *mgcv* package, and we illustrate this idea in our data application, though we leave rigorous examination of this idea for future work.

S-4 Additional Simulation Results

S-4.1 Data Driven Simulation

To investigate the differences in estimating the mean and principal components between vbFPCA and fast-GFPCA approaches we conducted a short simulation study. First the parameters of model (2) were estimated for $K = 4$ eigenfunctions using the non-overlapping fast-GFPCA model using a window size of $w = 6$. We generated five data sets with $N = 4000$ participants, each observed on the same grid as the original data ($J = 1440$). Since we did not find parameter estimates differed in the fast-GFPCA approach by bin width and overlapping versus non-overlapping windows, we only estimate the fast-GFPCA model with $w = 6$ and non-overlapping windows.

Figure S3 plots the results of this small simulation study and is structured the same as Figure 2, with the estimated population mean function $\hat{\mu}(s)$ and eigenfunctions $\hat{\phi}_k(s)$ plotted in the first and second through fifth rows, respectively. The true mean and eigenfunctions are plotted as solid black lines, while the estimated values across the 5 simulated datasets are plotted as semi-transparent red (vbFPCA) and blue (fast-GFPCA) models,

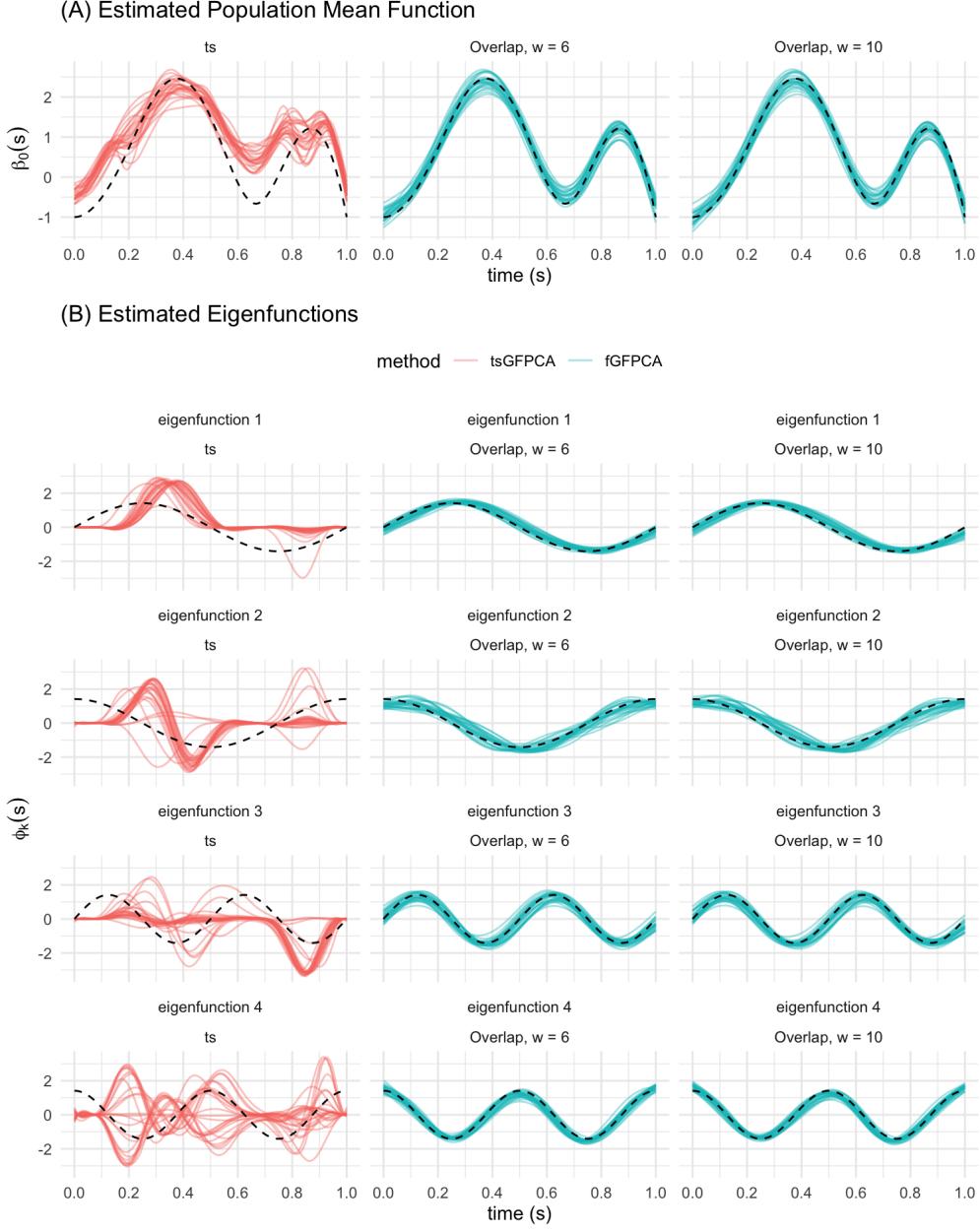


Figure S2: The estimated population mean function ($\hat{\beta}_0(s)$) and the first four estimated eigenfunctions ($\hat{\phi}_k(s)$) from 25 simulated Poisson functional datasets with $N = 100$ subjects and $J = 100$ time points. Model estimates are presented as red lines (fastGFPCA) or blue lines (fastGFPCA). fastGFPCA estimates compared with two fastGFPCA models. The fastGFPCA models (blue columns) are estimated with overlapping windows with bin sizes $w_l = 6$ and $w_l = 10$.

respectively. We see that both the population mean function and eigenfunction estimates are biased in the vbFPCA approach and the bias aligns remarkably well with the estimates we obtained from the real data. In contrast, fast-GFPCA estimates show minimal bias. As

expected with the relatively large sample size $N = 4000$), estimates of both the population mean and eigenfunctions are extremely consistent across simulated datasets. These findings suggest that the vbFPCA approach may indeed be providing biased parameter estimates in our data application.

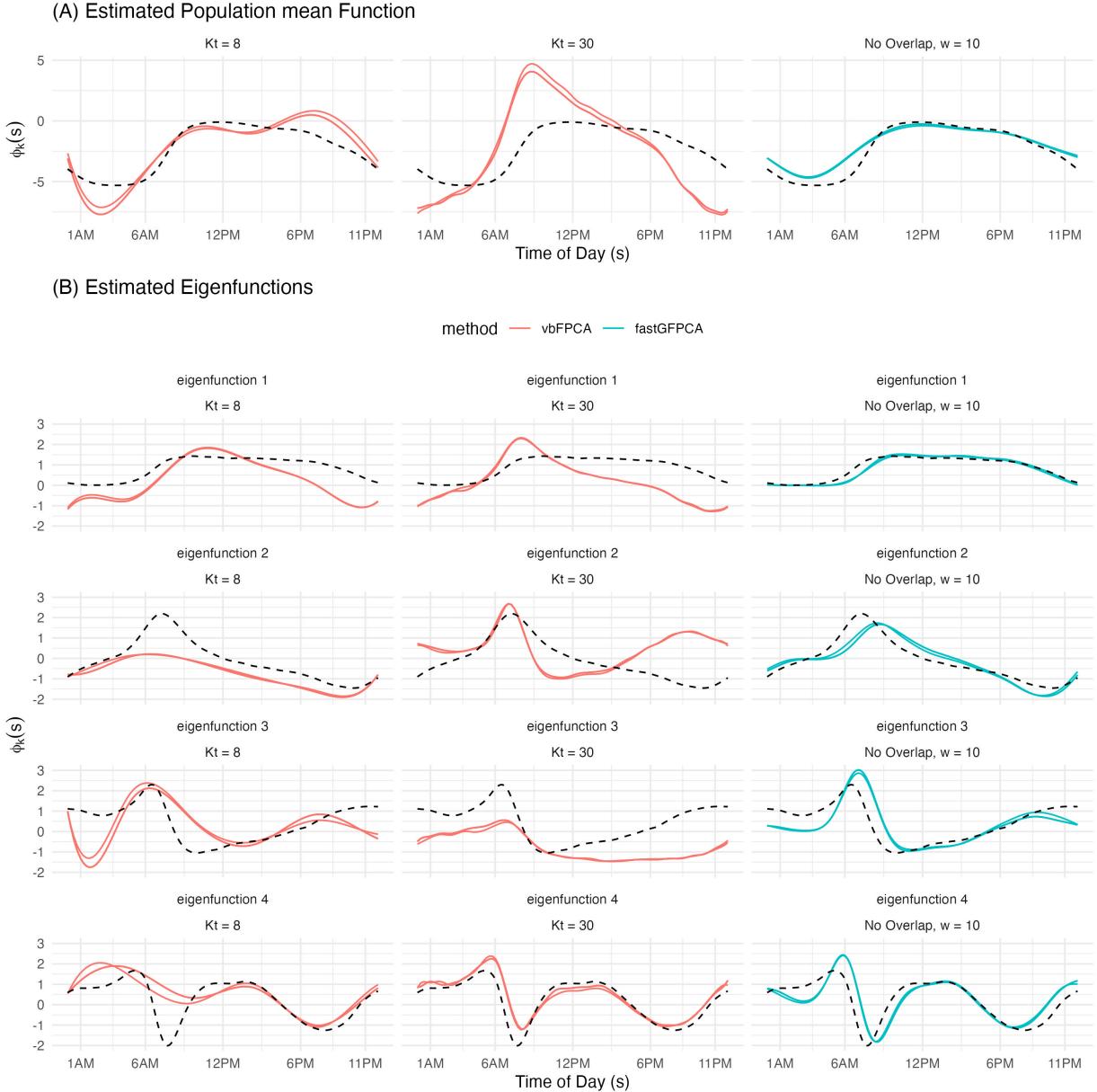


Figure S3: Figure showing the estimated population mean function ($\hat{\mu}(s)$) and the first four estimated eigenfunctions ($\hat{\phi}_k(s)$) from 5 simulated datasets based on the NHANES fast-GFPCA results. Model estimates are presented as red lines (vbFPCA) or blue lines (fast-GFPCA). vbFPCA estimates from models using either $K_t=8$ (left column) or $K_t=30$ (middle column) basis functions are presented and compared with fast-GFPCA estimated using non-overlapping windows of size $w_l = 10$.

S-4.2 Simulation results for non-periodic functional data

While our paper focuses mainly on the simulation results for data simulated with periodic endpoints, we also implemented simulation scenarios with non-periodic curves. The results of these scenarios are summarized in Table S1 below.

Integrated Squared Error of Model Parameters for Non-Periodic Curves								
Estimate	N	J	Fast-GFPCA			vbFPCA		tsGFPCA K = 30
			Non-Overlapping Bins			K = 8	K = 30	
$\psi(s)$	100	100	0.95	0.98	-	0.54	1.28	0.53
		500	0.55	0.49	1.41	0.31	0.52	-
		2000	0.63	0.5	0.4	0.52	6.85	-
	500	100	0.44	0.48	-	0.16	0.28	0.22
		500	0.24	0.18	1.08	0.09	0.14	-
		2000	0.21	0.14	0.14	0.36	6.08	-
	1000	100	0.34	0.42	-	0.1	0.17	-
		500	0.14	0.12	0.97	0.05	0.09	-
		2000	0.1	0.06	0.08	0.35	5.92	-
$\beta_0(s)$	100	100	11.82	13.57	-	16.29	28.51	10.09
		500	12.94	13.07	12.55	13.16	15.71	-
		2000	11.88	10.74	13.2	12.84	13.08	-
	500	100	1.68	1.66	-	3.71	6	2.24
		500	1.69	1.63	1.59	3.29	3.87	-
		2000	1.96	1.93	1.87	3.33	3.22	-
	1000	100	0.73	0.78	-	1.89	3.09	-
		500	0.77	0.76	0.71	1.5	1.77	-
		2000	0.83	0.74	0.7	2.21	1.76	-
$\eta_i(s)$	100	100	2.31	2.32	-	2.44	3.14	2.4
		500	0.56	0.56	0.58	0.58	0.73	-
		2000	0.16	0.16	0.17	0.15	0.19	-
	500	100	2.29	2.3	-	2.3	2.42	2.32
		500	0.59	0.59	0.61	0.55	0.58	-
		2000	0.16	0.16	0.17	0.15	0.15	-
	1000	100	2.3	2.31	-	2.29	2.34	-
		500	0.57	0.58	0.6	0.55	0.56	-
		2000	0.15	0.15	0.16	0.15	0.15	-

Table S1: Mean integrated squared error (MISE) for $\psi(s)$ and $\hat{\eta}_i(s)$ and integrated squared error (ISE) for $\beta_0(s)$ across methods and simulation scenarios. In each row the method(s) with the lowest MISE for that simulation scenario is in **bold**. All data summarized in this table were simulated with population mean $\beta_0(s) = 0$ and with non-periodic eigenfunctions. An “-” indicates that that model was not evaluated for a given simulation scenario.