Check for updates

Received: 6 November 2020 | Accepted: 2 May 2022

DOI: 10.1111/biom.13696

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

A Bayesian functional data model for surveys collected under informative sampling with application to mortality estimation using NHANES

Paul A. Parker[1] | **Scott H. Holan**[2,3]

[1]Department of Statistics, University of California Santa Cruz, Santa Cruz, California, USA

[2]Department of Statistics, University of Missouri, Columbia, Missouri, USA

[3]Research and Methodology Directorate, U.S. Census Bureau, Washington, DC, USA

Correspondence
Paul A. Parker, Department of Statistics, University of California Santa Cruz, 1156 High St, Santa Cruz, CA, 95064.
Email: paulparker@ucsc.edu

Funding information
National Science Foundation, Grant/Award Number: SES-1853096

Abstract

Functional data are often extremely high-dimensional and exhibit strong dependence structures but can often prove valuable for both prediction and inference. The literature on functional data analysis is well developed; however, there has been very little work involving functional data in complex survey settings. Motivated by physical activity monitor data from the National Health and Nutrition Examination Survey (NHANES), we develop a Bayesian model for functional covariates that can properly account for the survey design. Our approach is intended for non-Gaussian data and can be applied in multivariate settings. In addition, we make use of a variety of Bayesian modeling techniques to ensure that the model is fit in a computationally efficient manner. We illustrate the value of our approach through two simulation studies as well as an example of mortality estimation using NHANES data.

KEYWORDS

functional data analysis, horseshoe prior, National Health and Nutrition Examination Survey (NHANES), pólya-Gamma, pseudo-likelihood

1 | INTRODUCTION

The use of functional data as either a response or covariate has seen wide usage in recent years. Applications that utilize functional data include longitudinal data analysis (Yao *et al.*, 2005), ecology (Yang *et al.*, 2013), small area estimation (Porter *et al.*, 2014), as well as many others. However, models for functional data typically assume a sample that is representative of the population, and thus are not directly applicable to many survey data sets, especially under informative sampling. For example, the National Health and Nutrition Examination Survey (NHANES) contains functional data in the form of activity monitor curves, yet the survey design is complex leading to a sample that is not representative of the population.

There is a breadth of literature around functional data analysis (FDA). A seminal work in FDA is that of Ramsay and Silverman (2005), while Kokoszka and Reimherr (2017) provide a more recent treatment of the topic. The literature on FDA for survey data is quite sparse. Savitsky (2016) considers the case of functional responses under informative sampling. They treat the response as a Gaussian process to handle functional dependence while simultaneously using a weighted Bayesian pseudo-likelihood (BPL) to account for informative sampling. One drawback of this approach is that computation under the Gaussian process formulation can become prohibitively difficult in high-dimensional settings.

More recently, Leroux *et al.* (2019) explore scalar on function regression to predict 5-year mortality rate based on NHANES physical activity covariates. Their approach

 © 2022 The International Biometric Society.

employs existing software packages that use the survey weights to construct appropriate point estimates but are unable to give correct estimates of uncertainty based on the sample design. The authors state that a resampling procedure may be used to give appropriate standard errors, but the approach is beyond the scope of the paper. Ultimately, their use of scalar on function regression was more exploratory and not intended to fully account for the survey design.

In this work, we develop a Bayesian model for scalar on function regression of survey data under informative sampling. Through the use of a BPL (Savitsky and Toth, 2016), we are able to give appropriate measures of uncertainty. In addition, we use data augmentation to ensure that the model can be fit in an efficient manner via Gibbs sampling. We also provide an extension to Multinomial response data, which allows for certain multivariate problems to fit into our framework. Similar to Leroux *et al.* (2019), we are primarily motivated by the topic of mortality estimation with NHANES physical activity covariates, though we note that this methodology is generally applicable to many types of functional survey data found in practice. The remainder of this work is outlined as follows. In Section 2, we describe our methodology along with necessary background material. Section 3 outlines the motivating NHANES data set. In Section 4, we conduct a synthetic data simulation as well as an empirical simulation study that utilizes the public-use NHANES activity monitor data. We also present a data analysis involving public-use NHANES data in Section 5. Finally, we provide concluding remarks and discussion in Section 6.

2 | METHODOLOGY

2.1 | Informative sampling

In many survey data settings, there is dependence between a unit's probability of selection and the response of interest. This is termed *informative sampling* and is known to introduce bias into the model when ignored. Thus, in survey data settings, it is important to account for the survey design in some manner in order to eliminate or reduce this bias. In other words, complex sample designs can lead to samples that are unrepresentative of the population and, thus, the sample model should be adjusted in some way to account for this.

Parker *et al.* (2019) provide an overview of various methods to account for informative sampling. Of primary interest is the pseudo-likelihood (PL) method introduced by Skinner (1989) and Binder (1983). This approach adjusts the likelihood function by exponentially weighting each unit's likelihood contribution by the corresponding survey weight (ie, the inverse of the selection probability), $\prod_{i\in S} f(y_i \mid \theta)^{w_i}$, where S indicates the sample and y_i represents the response value for unit i with survey weight w_i. In a frequentist setting, this PL can be maximized to give a point estimate for θ, however more complex procedures are necessary to give appropriate estimates of uncertainty.

Savitsky and Toth (2016) show that a PL may also be used in a Bayesian framework. In particular, they show that under informative sampling, the use of a PL along with a prior specification leads to a pseudo-posterior distribution,

$$\widehat{\pi}(\theta|\mathbf{y},\widetilde{\mathbf{w}}) \propto \left\{\prod_{i\in S} f(y_i|\theta)^{\widetilde{w}_i}\right\}\pi(\theta), \tag{1}$$

that converges to the population posterior distribution. In this scenario, it is important to scale the weights to sum to the sample size in order to attain the appropriate estimates of uncertainty. For intuition, here, consider that an unweighted model is the same as the case where all weights are equal to one. Thus, the weights sum to the sample size. However, in practice, survey weights sum to roughly the population size. Inclusion of these weights directly would result in an effective sample size of roughly the population size, and extreme underestimation of the uncertainty. It is important to note that the weights in this context are fixed and known from the survey design. The case where the weights are random results in a more complex problem that may pose deficiencies in terms of Bayesian inference (eg, see Rubin, 1981). These scaled weights are represented by $\widetilde{w}_i$. This formulation applies generally to Bayesian models and is the approach we use herein to account for informative sampling.

For (1), the main result (asymptotics) of Savitsky and Toth (2016) relies on several technical assumptions. Their condition (A5) restricts the result to sampling designs where the dependence among lowest-level sampled units diminishes as ν (a sequence of finite populations) increases to infinity. For example, Savitsky and Toth (2016) point out that a two-stage cluster within strata sampling design, would meet this condition if the number of population units nested within each cluster (from which the sample is drawn) increases in the limit of ν. They also note that this would be the case in a survey of units within each cluster, if the cluster domains are defined geographically and grow in area as ν increases.

2.2 | Non-Gaussian data

Modeling non-Gaussian data types in a Bayesian setting can be computationally burdensome, especially while accounting for informative sampling. Parker *et al.* (2020)

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

utilize a data augmentation approach to construct a flexible mixed model for Binomial and Multinomial data under informative sampling. Their model for Binomial data is given by

$$\begin{aligned} \boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S} \text{Bin}(Z_i|n_i,p_i)^{\widetilde{w}_i} \\ \text{logit}(p_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} \\ \boldsymbol{\eta}|\sigma_\eta^2 &\sim \text{N}_r\left(\boldsymbol{0}_r, \sigma_\eta^2\boldsymbol{I}_r\right) \\ \boldsymbol{\beta} &\sim \text{N}_q\left(\boldsymbol{0}_q, \sigma_\beta^2\boldsymbol{I}_q\right) \\ \sigma_\eta^2 &\sim \text{IG}(a,b) \\ \sigma_\beta, a, b &> 0, \end{aligned} \tag{2}$$

where Z_i represents the response value for unit i in the sample. In this case, $\boldsymbol{x}_i$ is a vector of fixed effects covariates and $\boldsymbol{\phi}_i$ represents a set of spatial basis functions.

In order to fit this model in a computationally efficient manner, Pólya-Gamma data augmentation is used. Specifically, letting $\text{PG}(\cdot,\cdot)$ represent a Pólya-Gamma distribution, Polson *et al.* (2013) show that

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b}e^{\kappa\psi}\int_0^\infty e^{-\omega\psi^2/2}p(\omega)d\omega, \tag{3}$$

where $\kappa = a - b/2$ and $p(\omega)$ is a $\text{PG}(b,0)$ density. They further show that $(\omega|\psi) \sim \text{PG}(b,\psi)$. The PL in (2) can be written

$$\prod_{i\in S}\left\{\frac{(e^{\psi_i})^{Z_i}}{(1+e^{\psi_i})^{n_i}}\right\}^{\widetilde{w}_i} = \prod_{i\in S}\frac{(e^{\psi_i})^{Z_i^*}}{(1+e^{\psi_i})^{n_i^*}}, \tag{4}$$

where $\psi_i = \text{logit}(p_i)$, $Z_i^* = Z_i \times \widetilde{w}_i$, and $n_i^* = n_i \times \widetilde{w}_i$. This allows for data augmentation of a latent Pólya-Gamma random variable that leads to conjugate Normal priors on the regression parameters.

The Binomial model in (2) can also be extended to Multinomial or Categorical data. Following Linderman *et al.* (2015), the Multinomial distribution with C categories can be rewritten as

$$\text{Multinomial}(\boldsymbol{Z}|n,\boldsymbol{p}) = \prod_{c=1}^{C-1}\text{Bin}(Z_c|n_c,\widetilde{p}_c), \tag{5}$$

where $n_c = n - \sum_{j<c} Z_j$ and $\widetilde{p}_c = \frac{p_c}{1-\sum_{j<c}p_j}$ for $c = 2,\ldots,C$. In this light, a series of $C-1$ Binomial models may be fit to estimate the parameters for a Multinomial data model.

The modeling framework of Parker *et al.* (2020) is useful for fitting Binomial data under informative sampling, such as the NHANES mortality data of interest; however, the approach must be extended in order to consider functional covariates.

2.3 | Functional covariates

Consider the case where we have J functional covariates and $\kappa_{ij}(t), t \in \mathcal{T}$ denotes the jth functional covariate ($j = 1,\ldots,J$) for unit i at time t. In our case, the domain is time, though other domains may be appropriate depending on the type of functional data. Then, (2) can be extended for functional covariates by letting

$$\text{logit}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{j=1}^{J}\int_{\mathcal{T}}\eta_j(t)\kappa_{ij}(t)dt, \tag{6}$$

where $\eta_j(t)$ is a functional regression parameter associated with functional covariate j. In what follows, we will assume $J = 1$ (and drop the subscript j), as is the case in our example, but we note that the approach is still applicable for $J > 1$.

In order to reduce the dimension of the problem, we can use a basis expansion representation. In particular, let $\{\phi_k(t) : k = 1,2,\ldots\}$ be a complete orthonormal basis of the domain $\mathcal{T}$. Then, we can represent the functional covariate as

$$\kappa_i(t) = \sum_{k=1}^{\infty}\xi_i(k)\phi_k(t) \tag{7}$$

and

$$\eta(t) = \sum_{k=1}^{\infty}b(k)\phi_k(t), \tag{8}$$

where $\xi_i(k)$ and $b(k)$ are the expansion coefficients for $\kappa_i(\cdot)$ and $\eta(\cdot)$, respectively. Now, appealing to orthonormality,

$$\text{logit}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \int_{\mathcal{T}}\eta(t)\kappa_i(t)dt = \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{k=1}^{\infty}b(k)\xi_i(k). \tag{9}$$

Note that any orthonormal basis may be used here, though we use functional principal components selected through the fast covariance estimation (FAST) approach (Xiao *et al.*, 2016). This is easily implemented via the use of the `refund` package in `R` (Goldsmith *et al.*, 2019).

In practice, the summation in (9) is truncated to K. For our purposes, we truncate the summation (ie, choose K) such that the retained components explain 90% of the

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

variation in the functional data. This results in a finite, though potentially large, number of basis functions. Furthermore, any given basis function may not necessarily be related to the response. Thus, we require some form of variable selection and shrinkage estimator.

2.4 | Horseshoe prior

In order to provide shrinkage to our functional regression coefficients, we utilize the Horseshoe prior introduced by Carvalho *et al.* (2010). Although many other methods of Bayesian variable selection exist, this has the advantage of being fully specified, without requiring hyperparameter selection, as well as providing minimal shrinkage to strong signals while still providing a high degree of shrinkage for noise. To implement the Horseshoe prior for (9), we use the following:

$$\begin{aligned} b(k)|\lambda_k, \tau &\overset{ind}{\sim} \mathrm{N}(0, \lambda_k^2\tau^2), k = 1, \dots, K \\ \lambda_k &\overset{ind}{\sim} \mathrm{C}^+(0,1) \\ \tau &\sim \mathrm{C}^+(0,1), \end{aligned} \tag{10}$$

where $\mathrm{C}^+(\cdot,\cdot)$ represents the Cauchy density truncated below at zero. This prior is considered a global–local shrinkage approach. This can be seen by recognizing that τ applies to all regression parameters and determines the overall level of shrinkage, whereas λ_k is local and applies to a specific coefficient. In this way, coefficients corresponding to noise can attain a higher degree of shrinkage than those with strong signals. The motivation behind the use of a shrinkage prior for the basis coefficients is that there is no reason that the coefficients that explain the large-scale variation are necessarily the ones associated with the response. By using a shrinkage prior, the important basis coefficients are learned in a semisupervised fashion. A similar strategy has been successfully implemented in Holan *et al.* (2010), among others.

The half-Cauchy priors used in the Horseshoe are not conjugate. However, Makalic and Schmidt (2015) use a data augmentation approach to allow for Gibbs sampling within the Horseshoe prior framework. In particular, they use a scale mixture representation of the half-Cauchy such that when $x \sim \mathrm{C}^+(0, A)$, then $x^2|a \sim \mathrm{IG}(1/2, 1/a)$ and $a \sim \mathrm{IG}(1/2, 1/A^2)$, where $\mathrm{IG}(a, b)$ represents the inverse Gamma distribution with shape parameter a and scale parameter b. This leads to an alternate formulation of the Horseshoe prior hierarchy that allows for straightforward Gibbs sampling.

2.5 | Functional data model under informative sampling

We now present our model for non-Gaussian data under informative sampling with functional covariates, which makes use of the modeling elements discussed so far:

$$\begin{aligned} \boldsymbol{Z}|\boldsymbol{\beta}, \boldsymbol{\eta} &\propto \prod_{i \in S} \mathrm{Bin}(Z_i|n_i, p_i)^{\widetilde{w}_i} \\ \mathrm{logit}(p_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{k=1}^{K} b(k)\xi_i(k) \\ \boldsymbol{\beta} &\sim \mathrm{N}_q(\mathbf{0}_{\boldsymbol{q}}, \sigma_\beta^2 \boldsymbol{I}_q) \\ b(k)|\lambda_k, \tau &\overset{ind}{\sim} \mathrm{N}(0, \lambda_k^2\tau^2), k = 1, \dots, K \\ \lambda_k^2|\nu_k &\overset{ind}{\sim} \mathrm{IG}(1/2, 1/\nu_k) \\ \tau^2|\nu_\tau &\sim \mathrm{IG}(1/2, 1/\nu_\tau) \\ \nu_1, \dots, \nu_K, \nu_\tau &\overset{ind}{\sim} \mathrm{IG}(1/2, 1) \end{aligned} \tag{11}$$

for $\sigma_\beta^2 > 0$. In this model, $\boldsymbol{x}_i$ represents a q-dimensional vector of scalar covariates and $\xi_i(k)$ represents the kth basis expansion coefficient for observation i.

This model makes use of a Bayesian PL to account for informative sampling, allowing for population level inference. We also make use of Pólya-Gamma data augmentation for efficient Gibbs sampling. If desired, prior information on the scalar covariates may be incorporated through the selection of σ_β^2, though we use a relatively diffuse prior by letting $\sigma_\beta^2 = 10$. The full conditional distributions are given in Web Appendix A. As an alternative, one could use a generic Bayesian model fitting software such as Stan (Carpenter *et al.*, 2017) for small to moderate sample sizes, however, we have found Gibbs sampling to be more efficient as the dimensionality of the problem increases.

It is straightforward to implement the model in the Multinomial data setting through the use of a stick-breaking representation, as discussed in Section 2.2. It would also be straightforward to use a Gaussian PL in place of the Binomial one given here, as conjugacy would be retained.

3 | NHANES DATA DESCRIPTION

The NHANES is a survey conducted by the National Center for Health Statistics that utilizes a complex survey

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

design to collect health and nutrition data in the United States. Of primary interest to us is the physical activity monitor (PAM) data collected during the 2003–2004 and 2005–2006 samples. Along with this, we are interested in mortality as a response value.

NHANES provides microdata to the public, however, a substantial amount of data processing is required to utilize the data for inference. Leroux *et al.* (2019) provide very helpful exposition on processing the data as well as the `rnhanesdata` package in `R` for doing so. All analyses in this work were conducted using data that were prepared and processed in the same manner as Leroux *et al.* (2019).

In particular, we use the NHANES samples from 2003 to 2004 and 2005 to 2006, as these contain the PAM data of interest. For each minute during a seven consecutive day period, the data contain an activity intensity value for each subject. Not all subjects were in compliance, resulting in variation in wear-time between subjects. Thus, all subjects that had less than 3 days of 10 h or greater wear-time were dropped. In addition, subjects outside of the age range 50–85, or who were missing mortality or age data, were also dropped. The resulting sample size was 3208.

The PAM activity measurements were transformed using $f(x) = \log(1 + x)$. The subjects in the sample had a varying numbers of days with activity data. To account for this, we use the PAM data averaged across days within subjects, resulting in a single 24-h curve for each subject.

The NHANES sample contains the required survey weights. Following Leroux *et al.* (2019), these are reweighted to account for missing data. More in-depth reweighting schemes may be desired in practice. However, discussion on reweighting procedures for missing data is beyond the scope of this work. Thus, for illustration, we do not consider this problem further.

4 | SIMULATION STUDIES

This section includes two different simulation studies. The first is a synthetic data simulation, based on prespecified and known functional coefficients, where a two-stage sample under an informative design is taken. The primary goal of the first simulation is to evaluate the quality of our uncertainty estimates around the functional coefficient.

Our second simulation is an empirical simulation study where we treat the existing NHANES sample as a population and take a further informative subsample. Simulating in this manner allows us to retain many of the characteristics of the NHANES data, while having a known truth to compare out of sample predictions to. Note that in this case, unlike the synthetic data simulation, the true functional coefficient is unknown.

4.1 | Synthetic data simulation

In order to assess the ability of our model to perform inference on a population level functional coefficient, we create a synthetic population constructed under a known functional coefficient. In this case, we consider both a very smooth function (Function A) as well as a more variable function (Function B). These are shown as dashed lines in Figure 1.

We create a population of 100,000 individuals, and assign each individual into one of 100 different clusters. The cluster probabilities are generated from a Dirichlet distribution with all concentration parameters set to one. This results in a population with uneven cluster memberships. To generate the individual functional covariates, we consider weighted combinations of the functional covariates observed in the NHANES sample. For each individual, we generate a set of weights by drawing from a Dirichlet distribution with concentration parameters set to 1/3208 (where the denominator is the number of observations in the NHANES data). This results in synthetic curves that place a large amount of weight on a few curves, and little weight on the remaining curves. This process results in curves that look quite similar to the ones observed in the NHANES data, without any two curves being identical. In addition to the functional covariate, we generate one categorical covariate with four categories and corresponding coefficients drawn from a standard normal distribution. Finally, using the synthetic covariates and regression coefficients, we generate a linear predictor for each individual. We apply the inverse logit transformation to the linear predictors in order to generate a mortality probability for each individual, and then generate responses by drawing Bernoulli random variables conditional on the probabilities.

After generating a synthetic population, we take an informative sample to be used for estimation. In this case, we take a two-stage sample. In the first stage, we sample 10 of the available clusters. Clusters are sampled with probability proportional to the number of individuals within the cluster. Thus, clusters with more membership are more likely to be included in the sample. Cluster inclusion probabilities may be denoted as p_c. In the second stage, for each cluster chosen in stage one, we use Poisson sampling (Brewer *et al.*, 1984) to retain an expected sample size of 10% of the cluster size. The probability of selection of unit i in cluster c, conditional on cluster c being in the first-stage sample, is denoted as $p_{i|c}$. For true function A, we set this probability to be proportional to $\exp\{3.5 \times I(Z_i = 1) + 2 \times I(X_{i,cat} = 1) + 0.5 \times I(X_{i,cat} = 2)\}$, where $I(Z_i = 1)$ indicates that the response for individual i is 1. Similarly, $I(X_{i,cat} = k)$ indicates that the categorical covariate for

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

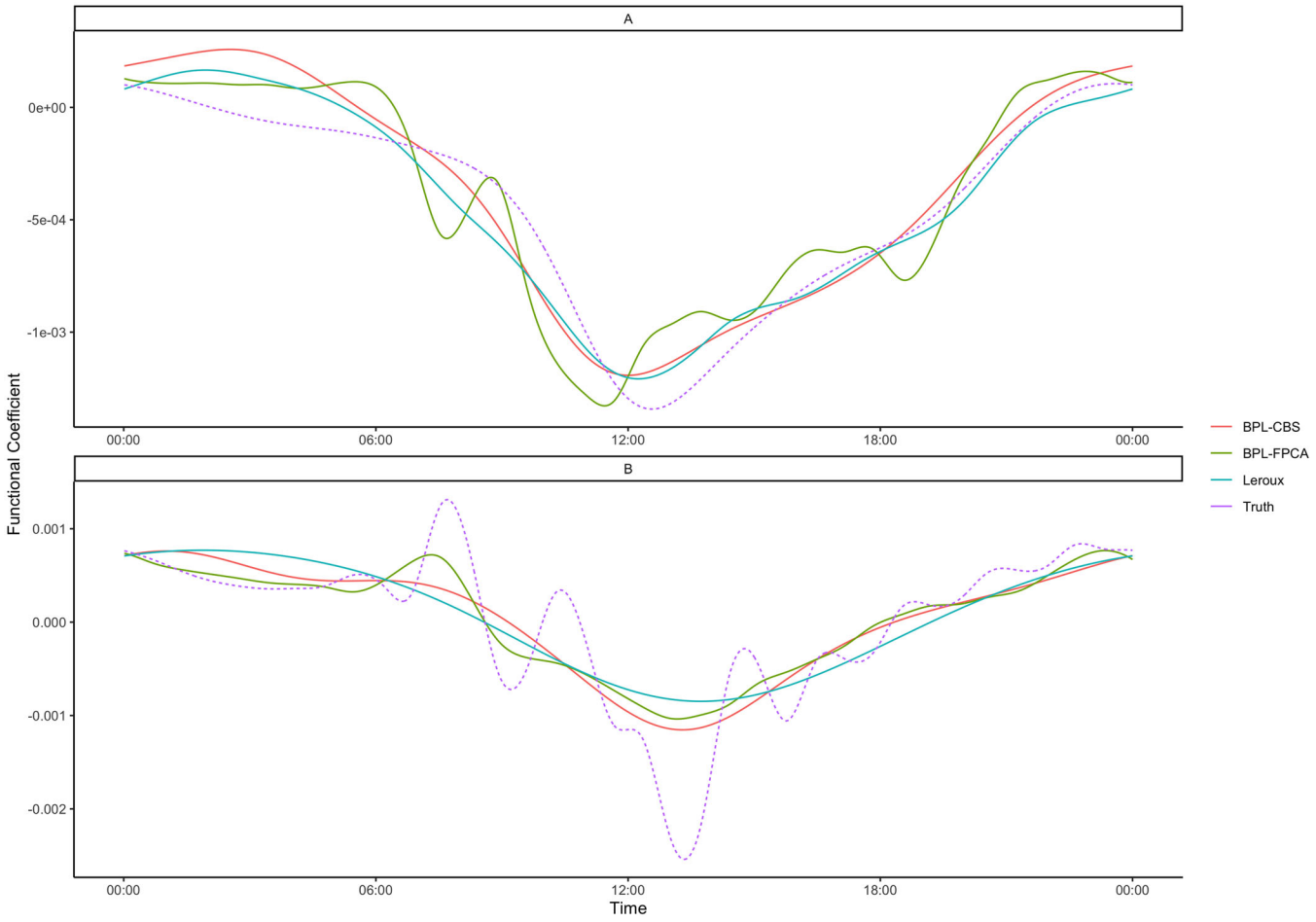

FIGURE 1 Comparison of true functional coefficient to estimated functional coefficient (averaged across simulated data sets). The true curves are shown as dashed lines. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

unit i is equal to k. For true function B, we use probability proportional to $\exp\{1.5 \times I(Z_i = 1) + 0.7 \times I(X_{i,cat} = 1) + 0.5 \times I(X_{i,cat} = 2)\}$. The use of categorical variables in the second stage makes this stage similar to a stratified random sample. Inclusion of the response in this second stage forces an informative design. In practice, informativeness is often an unintended consequence of a given sample design, rather than a feature. Finally, after sampling both stages, the marginal inclusion probabilities can be denoted as $p_i = p_{i|c} \times p_c$.

Using this two-stage sample design, we take 200 samples each from the synthetic populations under functional coefficients A and B. For each sample, we fit our Bayesian PL model using functional principal components basis functions (BPL-FPCA), and, similar to Leroux *et al.* (2019), cyclic B-spline basis functions of order 30 (BPL-CBS). We also compare to the frequentist model fit by Leroux *et al.* (2019), which uses cyclic B-spline basis functions of order 30 as well. For reference, the BPL models took roughly 6 min each for computation.

TABLE 1 Synthetic data simulation results comparing mean absolute error (MAE), absolute bias, and 95% credible/confidence interval coverage rate

Function	Model	MAE ($\times 10^4$)	Abs. bias ($\times 10^5$)	Coverage
A	BPL-FPCA	12.4	15.4	90.7%
	BPL-CBS	7.5	11.5	**93.4%**
	Leroux	6.4	9.7	62.1%
B	BPL-FPCA	6.9	25.6	**96.2%**
	BPL-CBS	5.2	28.9	92.4%
	Leroux	4.6	33.8	73.2%

Note: All results are averaged pointwise along the functional coefficient and across all simulated data sets.

For each sample, we are able to compare the functional coefficient estimates, along with their corresponding uncertainty, to the truth. Table 1 provides a summary of these results. We present the mean absolute error (MAE), the absolute bias, and the 95% credible/confidence

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

interval coverage rates. Each metric is averaged across time (the domain of the functional coefficient) as well as across the sampled data sets. In general, the BPL-FPCA model seems to result in slightly worse point estimates in terms of MAE. Although the model used by Leroux *et al.* (2019) results in low MAE, particularly when the true function is smooth, it severely underestimates the uncertainty in both the smooth and unsmooth settings. In comparison, both Bayesian PL models result in coverage rates that are much closer to optimal (ie, the nominal level). In general, for this simulation, the BPL-CBS model seems to strike a strong balance between selecting good point estimates for the functional coefficient and quantifying the uncertainty appropriately.

We also compare the true functional coefficients to the average coefficients across the sampled data sets for each model in Figure 1. For true function A, the models that use cyclic B-splines follow the truth quite well. The BPL-FPCA model, while able to capture the general trend, has additional "noise" that increases the MAE. In setting B, the cyclic B-spline models are able to capture the general trend, but unable to follow some of the peaks and valleys in the true function. In contrast, the BPL-FPCA model is better able to capture some of these high and low points (though not perfectly).

The simulation results do not give indication that there is a clear favorite in terms of which set of basis functions to use. Rather, it most likely depends on the properties of the data as well as the goals of the analysis. Our methodology is not beholden to any given set of basis functions. It is straightforward to swap the FPCA and cyclic B-spline bases for others and we encourage practitioners to explore the effects of these decisions.

4.2 | Empirical simulation

The goal of our empirical simulation is twofold. First, we want to confirm that the model is able to adjust for an informative sampling mechanism in order to allow for population level inference. Second, we want to assess whether or not the use of functional covariates, as given in the model, leads to improved estimates for units in the population.

To design such a simulation, we begin by treating the existing NHANES sample data as our population. This provides a baseline truth for which we can compare to. Next, we subsample from the NHANES data in an informative manner. Doing so, we are able to fit the model using the subsampled data and then compare to the population truth (ie, the original sample data). To take this subsample, we use probability proportional to size sampling via the Poisson method (Brewer *et al.*, 1984) with an expected sample size of 500. We construct the size variable as $s_i = \exp\{w_i^* + 2 * I(Z_i = 1)\}$ where w_i^* is the NHANES reported survey weight after scaling to have mean zero and variance 1, and $I(Z_i = 1)$ is an indicator that the ith respondent died within 5 years of the survey. Through this subsampling procedure, we obtain a new set of weights that are the inverse probabilities of selection.

The response data of interest is the binary indicator of 5-year mortality. We use age along with an intercept term as a scalar covariate and the PAM data as a functional covariate. After subsampling, we fit both the BPL-CBS and BPL-FPCA models as well as the model used by Leroux *et al.* (2019). We also fit a basic version of the Bayesian PL model that uses only the scalar covariate and disregards the functional data (BPL-S). In addition, we implement unweighted versions of the BPL-FPCA and BPL-S models (UW-FPCA and UW-S, respectively).

After fitting each of these models, we are able to make mortality predictions for the entire population. This results in a 5-year probability of mortality for each person in the population that we can compare to their actual mortality result. Because these outcomes are binary, we use binary cross-entropy (BCE) as a loss metric to compare rather than mean-squared error. This is calculated as

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\{Z_i \log(\hat{p}_i) + (1 - Z_i)\log(1 - \hat{p}_i)\}, \quad (12)$$

where $\hat{p}_i$ is the posterior mean probability of mortality for unit $i = 1, \ldots, N$ in the population with size N. Note that a lower value of BCE indicates a better model fit. We calculate the average BCE over the entire population, rather than the sample, as a measure of model fit and out of sample model prediction ability.

We repeat this subsampling and model fitting procedure 50 times, resulting in a distribution of BCE loss values under each model. We compare these distributions in Figure 2. It is immediately clear that the two unweighted models perform much worse than the weighted models. These unweighted models do not account for the informative sample design and thus introduce a large amount of bias when making inference on the population. The weighted models are able to account for the sample design, and thus result in much lower values of BCE. In addition, the distributions of BCE under the BPL-CBS and BPL-FPCA models are shifted to the left of BPL-S, indicating that the functional covariate does aid in prediction of mortality for members of the population. Interestingly, the Leroux model performs slightly worse than the BPL-CBS and BPL-FPCA models, perhaps due to oversmoothing of the functional coefficient. These results indicate that for population level inference based on the full-sample data,

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

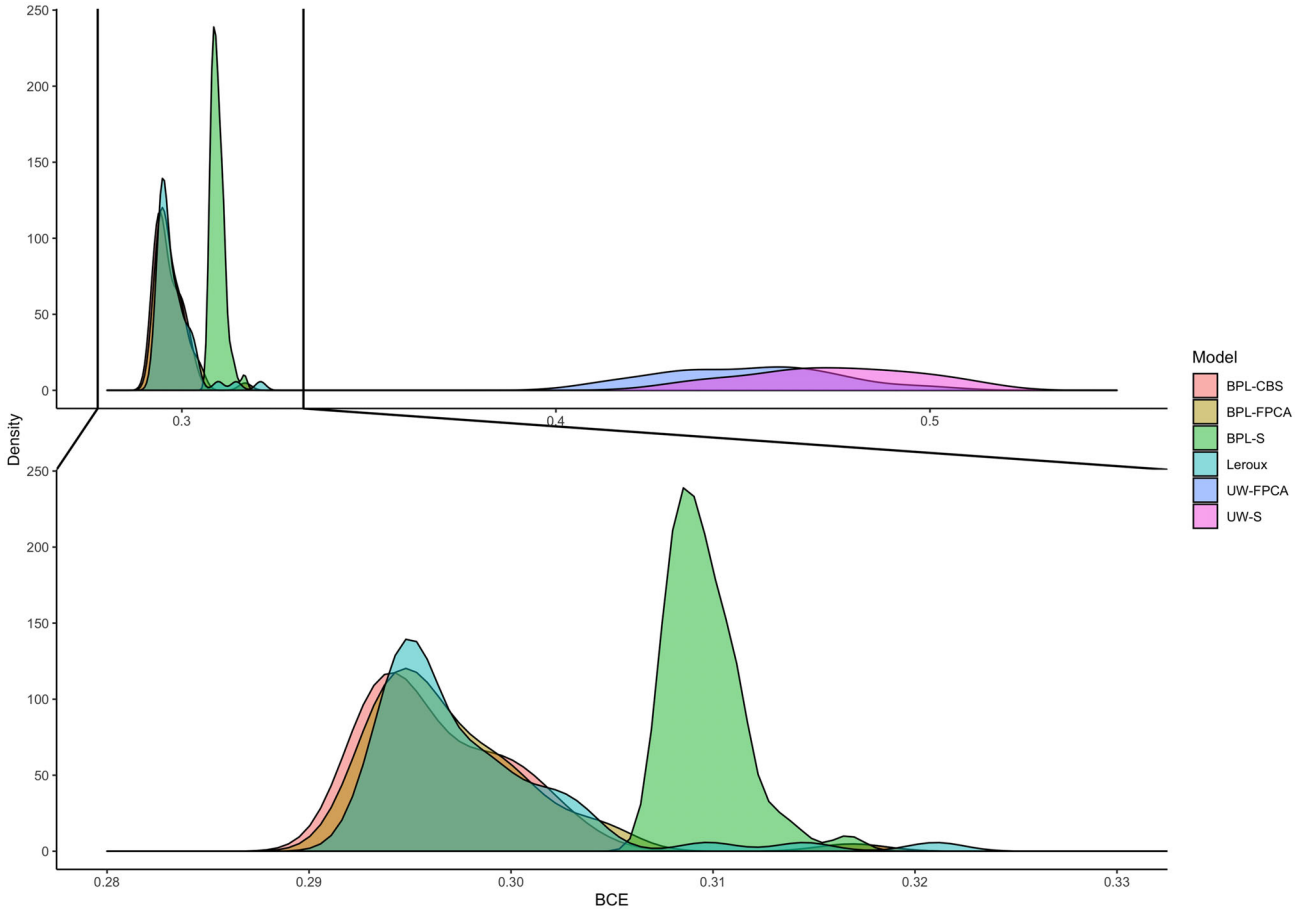

FIGURE 2 Simulation-based distribution of BCE values under each model. The lower subplot focuses on the four weighted models which have substantially lower BCE. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

we should use the model that utilizes functional covariates while also accounting for the survey design through a Bayesian PL.

5 | NHANES DATA ANALYSIS

5.1 | Five-year mortality estimate

Using our BPL-CBS, we now analyze the NHANES PAM data and its relationship with mortality. We use the same data set considered in the simulation study and outlined in Section 3, with a sample size of 3208. We treat the 5-year mortality indicator as our binary response, and use age, gender, body mass index (BMI), race, education level, as well as self-reported presence of a mobility problem, diabetes, coronary heart disease (CHD), congestive heart failure (CHF), cancer, and stroke as our scalar covariates in addition to the PAM data as a functional covariate. Note that these are the same covariates considered by Leroux *et al.* (2019).

We fit the model via Gibbs sampling with 5000 iterations and discard the first 1000 iterations as burn-in. Convergence was assessed via traceplots of the sample chains, where no lack of convergence was detected. Computation time was roughly 6.5 iterations per second on a standard laptop, or roughly 13 min total.

After fitting the model, we are able to make population level inference. We plot the posterior mean of the functional regression coefficient, $\eta(t)$, reconstructed through the estimated basis coefficients, along with a pointwise 95% credible interval in Figure 3. For the most part, $\eta(t)$ is estimated to be negative, as expected, indicating that increased levels of activity are associated with lower expected mortality rate. The primary time period where the credible interval does not contain zero is around 12 p.m. The model used by Leroux *et al.* (2019) underestimates the uncertainty, and correspondingly results in a much larger significant window than the one obtained here.

In addition to examination of the functional regression coefficient, we can also glean insight by examining how

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

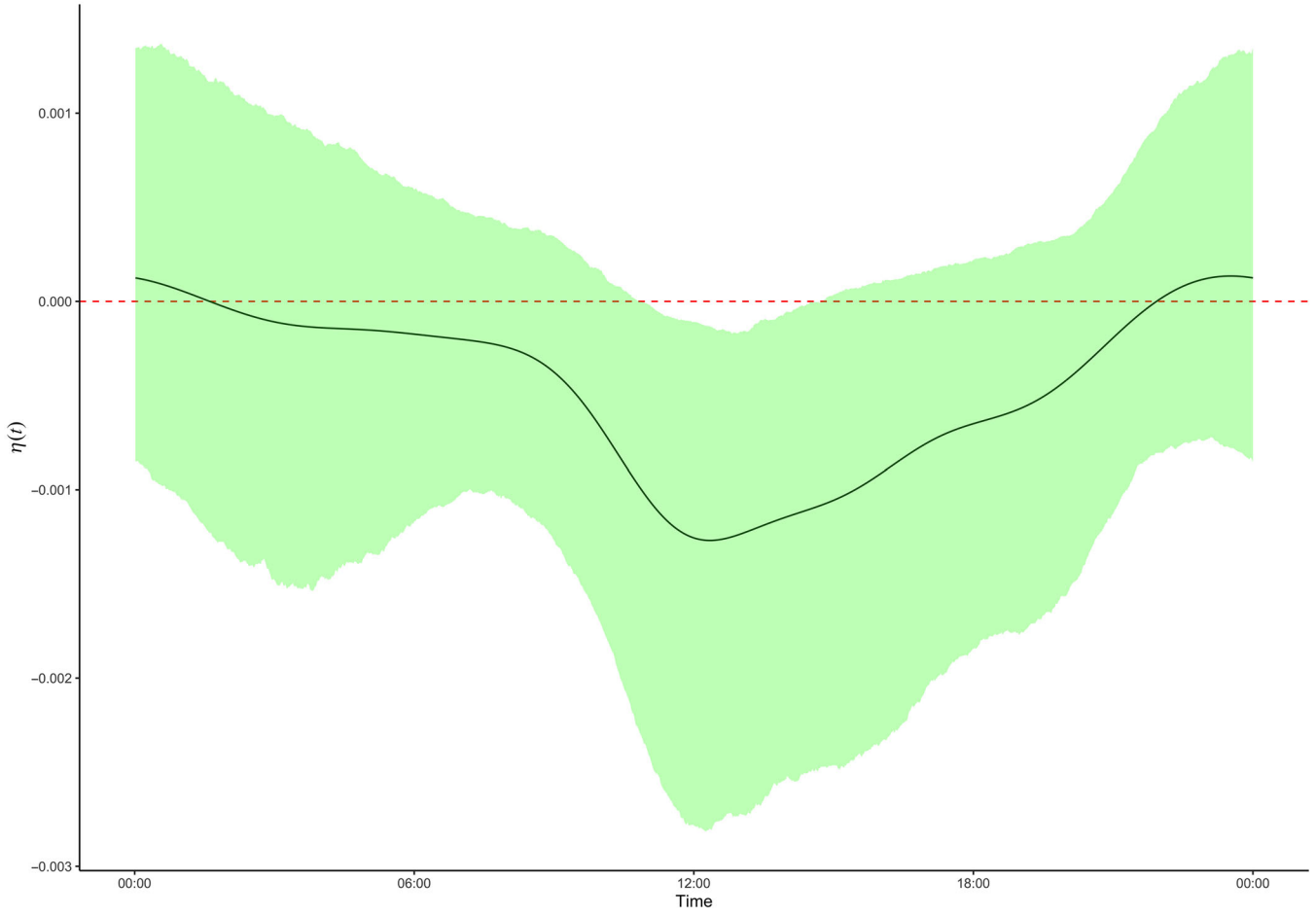

FIGURE 3 Estimate of the reconstructed PAM functional regression coefficient for 5-year mortality. The pointwise 95% credible interval for this estimate is shown as the shaded region. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

variation in activity level of individuals changes mortality estimates. In Figure 4, we plot activity curves for three individuals contained in the NHANES sample along with the accompanying posterior distribution of 5-year mortality rate. Individual A has a very low level of activity, resulting in a high expected mortality rate. However, there is also a great deal of uncertainty around this rate. Individuals B and C both have increasing level of overall activity, resulting in decreasing expected mortality rate. As the activity level increases, uncertainty around the mortality rate decreases.

5.2 | Joint mortality estimate

In addition to univariate estimation, our model allows for joint mortality estimation at multiple time scales through the Multinomial data model. In this case, we wish to make joint mortality estimates for years 1–5. To do so, we begin by assigning survey respondents into distinct categories: those who died within 1 year of the survey, those who died after 1 year but before 2 years, those who died after 2 years but before 3 years, those who died after 3 years but before 4 years, those who died after 4 years but before 5 years, and finally those who did not die before 5 years. Assigning groups in this way results in a Multinomial or Categorical data distribution with six categories. Thus, we are able to use the stick-breaking representation of the Multinomial distribution in order to fit $C - 1 = 5$ independent Binomial data models that allow us to make joint estimates of mortality at various time points.

We use the same individuals from our 5-year mortality example to examine the effects of activity level on multiyear mortality. Figure 5 plots the activity curves for these individuals alongside their posterior mean mortality rates for years 1–5. We also provide 95% credible intervals. Once again, we see that both expected mortality rate and uncertainty increase as activity level generally decreases. Because these estimates are multivariate, we can also see that decreased activity is associated with steeper marginal

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

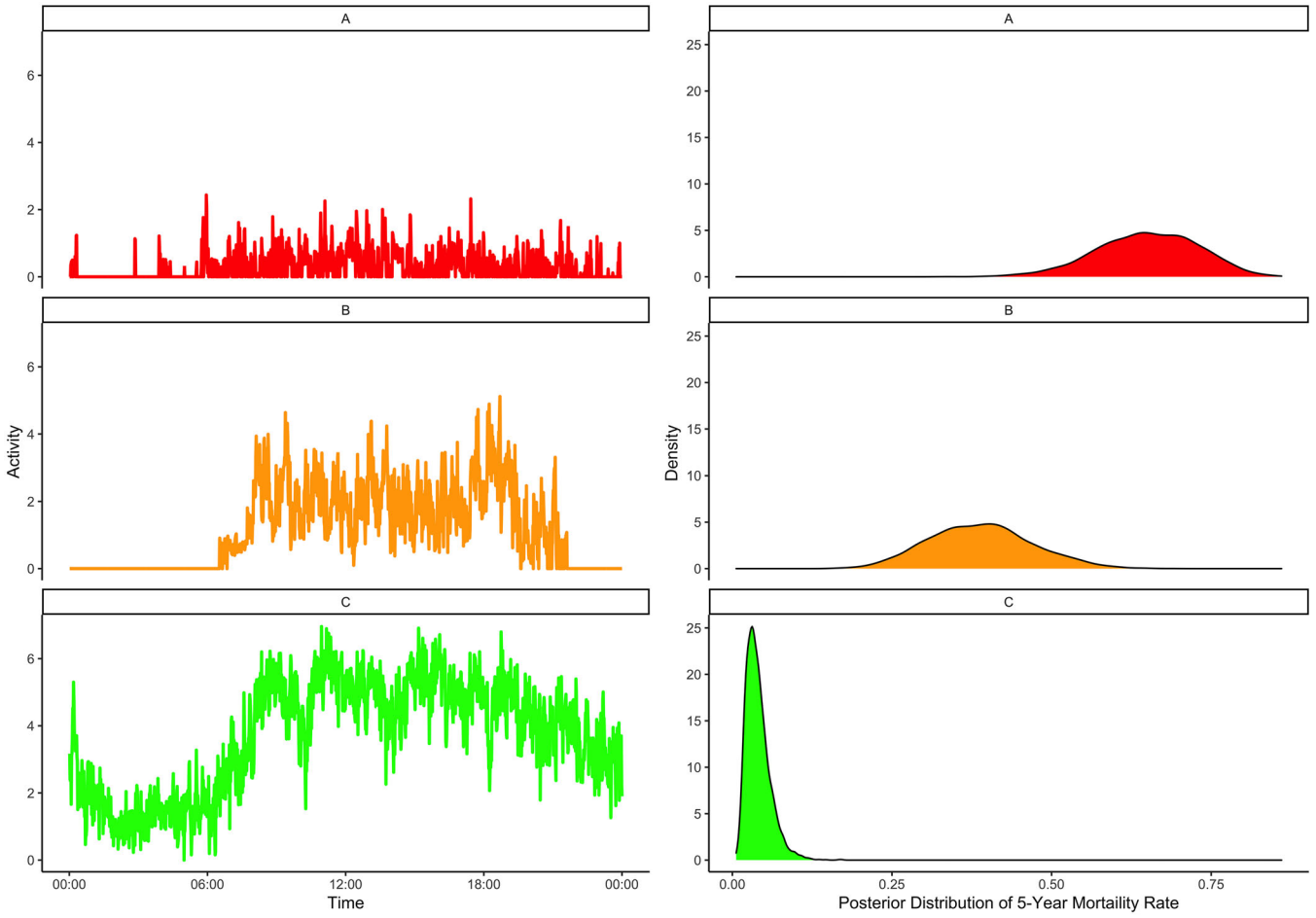

FIGURE 4 Activity curves for three individuals contained in the NHANES sample along with posterior distribution of 5-year mortality rate. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

increases in mortality for the near future than for years further away from the survey.

6 | DISCUSSION

In this work, we develop a Bayesian non-Gaussian data model for functional covariates under informative sampling. We rely on a PL approach to account for survey design which works in combination with Pólya-Gamma data augmentation to allow for conjugate full conditional distributions of the regression parameters. This method is designed for Binomial or Multinomial data, although extension to the Gaussian case is straightforward. Our approach uses a basis expansion representation of the functional covariates alongside the Horseshoe prior to provide regularization. Our approach allows for straightforward and efficient Gibbs sampling, which can be important in high-dimensional settings such as NHANES PAM data.

We conduct both a synthetic data simulation, as well as an empirical simulation study using NHANES data that show that our approach is able to reduce the bias attributable to informative sampling while also making use of the functional data to improve estimates for members of the population. Importantly, our method is able to give accurate uncertainty quantification as well as provide joint mortality estimates, improving over existing methodology. We also provide a full analysis of the NHANES data that allows us to make inference and prediction on the population. We conduct both a univariate analysis concerning 5-year mortality rate as well as a multivariate analysis concerning years 1–5 mortality rate.

Our methodology extends the literature on functional regression to the survey data setting. The approach is flexible in that users have a choice of data model and basis expansion and also allows for joint estimation of scalar regression coefficients. Currently, there is a limited amount of functional data collected under complex surveys, as analysis options are limited. It is our hope

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

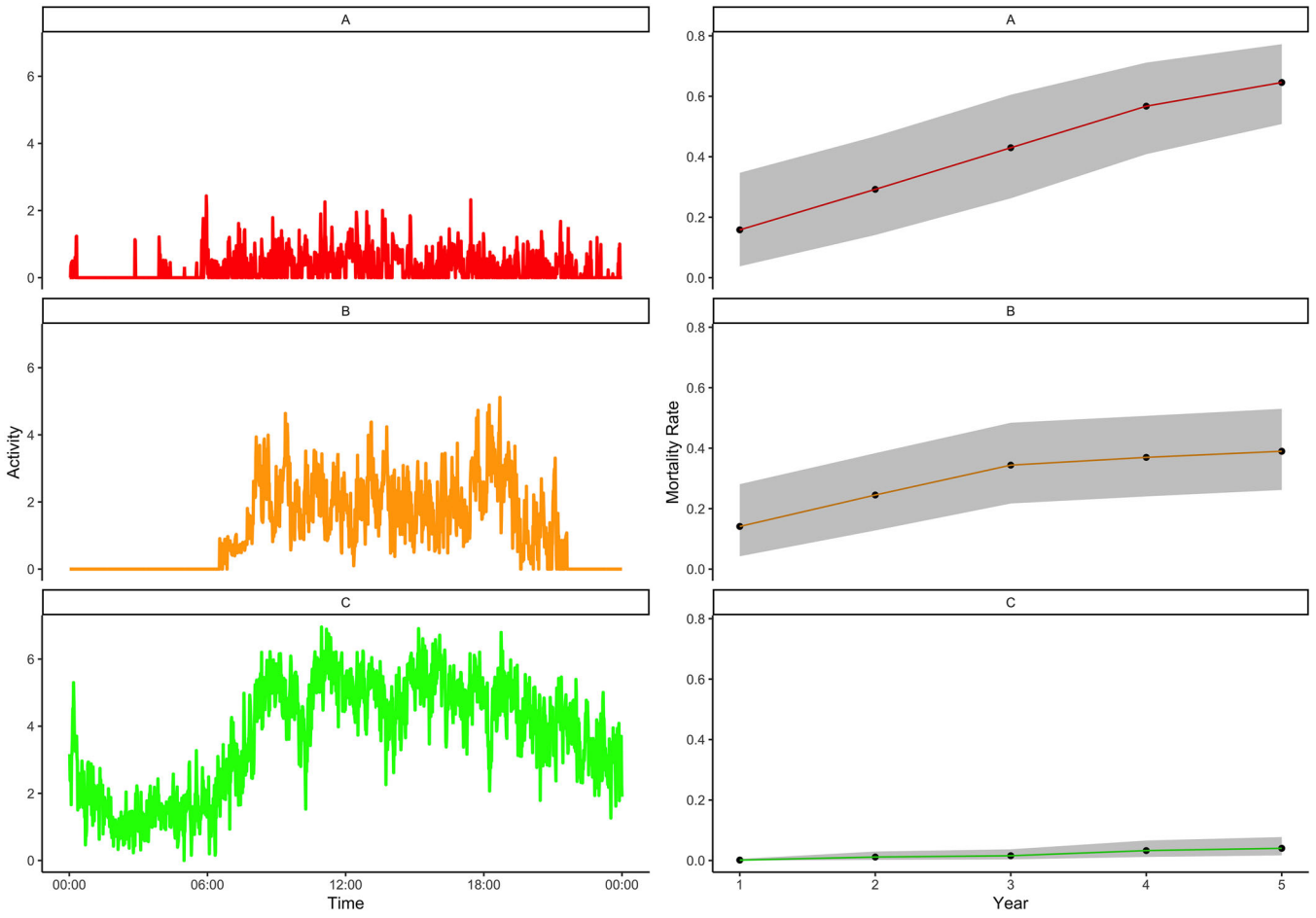

FIGURE 5 Activity curves for three individuals along with accompanying posterior mean mortality rate for years 1–5. The shaded region represents 95% credible intervals. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

that with the availability of this methodology, collection of functional data via surveys will become more widespread.

ACKNOWLEDGMENTS

We thank the Co-editor, the Associate Editor, and two anonymous referees for valuable comments that helped improve this paper. Support for this research through the Census Bureau Dissertation Fellowship program is gratefully acknowledged. This research was partially supported by the US National Science Foundation (NSF) under NSF Grant SES-1853096. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the NSF or US Census Bureau.

DATA AVAILABILITY STATEMENT

The data used in this paper are available in the public domain: https://wwwn.cdc.gov/Nchs/Nhanes/2003-2004/PAXRAW_C.ZIP/ https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/PAXRAW_D.ZIP https://wwwn.cdc.gov/Nchs/Nhanes/2003-2004/DEMO_C.XPT https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.XPT

ORCID

Paul A. Parker https://orcid.org/0000-0002-3206-6122

REFERENCES

Binder, D.A. (1983) On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279–292.

Brewer, K., Early, L. and Hanif, M. (1984) Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15–30.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M. et al. (2017) Stan: a probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010) The horseshoe estimator for sparse signals. *Biometrika*, 97, 465–480.

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J. et al. (2019) *refund: Regression with Functional Data*. R Package Version 0.1-21.

Holan, S.H., Wikle, C.K., Sullivan-Beckers, L.E. and Cocroft, R.B. (2010) Modeling complex phenotypes: generalized linear models using spectrogram predictors of animal communication signals. *Biometrics*, 66, 914–924.

Kokoszka, P. and Reimherr, M. (2017) *Introduction to Functional Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Leroux, A., Di, J., Smirnova, E., Mcguffey, E.J., Cao, Q., Bayatmokhtari, E. et al. (2019) Organizing and analyzing the activity data in NHANES. *Statistics in Biosciences*, 11, 262–287.

Linderman, S., Johnson, M.J. and Adams, R.P. (2015) Dependent multinomial models made easy: stick-breaking with the Pólya-Gamma augmentation. In *Advances in Neural Information Processing Systems*, pp. 3456–3464.

Makalic, E. and Schmidt, D.F. (2015) A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23, 179–182.

Parker, P.A., Holan, S.H. and Janicki, R. (2020) Computationally efficient Bayesian unit-level models for non-Gaussian data under informative sampling. *arXiv preprint arXiv:2009.05642*.

Parker, P.A., Janicki, R. and Holan, S.H. (2019) Unit level modeling of survey data for small area estimation under informative sampling: a comprehensive overview with extensions. *arXiv preprint arXiv:1908.10488*.

Polson, N.G., Scott, J.G. and Windle, J. (2013) Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108, 1339–1349.

Porter, A.T., Holan, S.H., Wikle, C.K. and Cressie, N. (2014) Spatial Fay–Herriot models for small area estimation with functional covariates. *Spatial Statistics*, 10, 27–42.

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

Rubin, D.B. (1981) The Bayesian bootstrap. *Annals of Statistics*, 9, 130–134.

Savitsky, T.D. (2016) Bayesian nonparametric mixture estimation for time-indexed functional data in R. *Journal of Statistical Software*, 72.

Savitsky, T.D. and Toth, D. (2016) Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10, 1677–1708.

Skinner, C.J. (1989) Domain means, regression and multivariate analysis. In: Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) *Analysis of Complex Surveys*. Chichester: Wiley, pp. 80–84.

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

Yang, W.-H., Wikle, C.K., Holan, S.H. and Wildhaber, M.L. (2013) Ecological prediction with nonlinear multivariate time-frequency functional data models. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 450–474.

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

SUPPORTING INFORMATION

Web Appendix A, referenced in Section 2, along with processed data and R code, are available with this paper at the Biometrics website on Wiley Online Library. In addition, the processed data and R code are available at https://github.com/paparker/survey_FDA.

How to cite this article: Parker, P.A. and Holan, S.H. (2023) A Bayesian functional data model for surveys collected under informative sampling with application to mortality estimation using NHANES. *Biometrics*79, 1397–1408.
https://doi.org/10.1111/biom.13696

15410420, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13696 by University Of Missouri Columbia, Wiley Online Library on [02/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License