Abstract

Functional Principal Component Analysis (FPCA) is a prominent tool to characterize variability and reduce dimension of longitudinal and functional datasets. Bayesian implementations of FPCA are advantageous because of their ability to propagate uncertainty in subsequent modeling. To ease computation, many modeling approaches rely on the restrictive assumption that functional principal components can be represented through a pre-specified basis. Under this assumption, inference is sensitive to the basis, and misspecification can lead to erroneous results. Alternatively, we develop a flexible Bayesian FPCA model using Relaxed Mutually Orthogonal (ReMO) processes. We define ReMO processes to enforce mutual orthogonality between principal components to ensure identifiability of model parameters. The joint distribution of ReMO processes is governed by a penalty parameter that determines the degree to which the processes are mutually orthogonal and is related to ease of posterior computation. In comparison to other methods, FPCA using ReMO processes provides a more flexible, computationally convenient approach that facilitates accurate propagation of uncertainty. We demonstrate our proposed model using extensive simulation experiments and in an application to study the effects of breastfeeding status, illness, and demographic factors on weight dynamics in early childhood. Code is available on GitHub: https://github.com/jamesmatuk/ReMO-FPCA.

Keywords: Bayes; Functional data analysis; Gaussian process; Orthogonality; Uncertainty quantification.
1 Introduction

Functional principal component analysis (FPCA) extends multivariate principal component analysis to functional and longitudinal data. This is achieved by acknowledging the infinite-dimensional nature of stochastic processes that underlie the observations and through enforcing smoothness in inferred functions (Ramsay & Silverman 2005, Ferraty & Vieu 2006). For functional and longitudinal data, functional principal components (FPCs) serve as an efficient orthogonal basis that captures dominant modes of observed variability, viewed as eigen-functions of an estimated covariance function. FPC scores are projections of the observations onto the FPC basis and serve as a low-dimensional representation of the data. FPCs themselves can be used to characterize and visualize the variability of observations, while FPC scores can be used directly in existing multivariate statistical and machine learning methods for classification, clustering, and regression tasks. Consequently, FPCA is a prominent tool for functional data analysis.

Some early approaches for FPCA, e.g., Castro et al. (1986), Rice & Silverman (1991), and Cardot (2000), assume that functional data are observed on the same dense grid of points. However, inference can fail when observations are recorded on sparse and irregular grids. Methods have been developed specifically to address this setting. In Yao et al. (2005), the authors pool information across observations to estimate mean and covariance functions based on kernel smoothing. Then, FPCA is carried out through eigen-decomposition of the estimated covariance function. In contrast, James et al. (2000) adapts mixed effects modeling approaches for functional data (Brumback & Rice 1998, Rice & Wu 2001) by specifying observations as linear combinations of FPCs. The FPCs are represented as expansions of a pre-specified basis. FPCA is carried out through inference on basis coefficients, where orthogonality on the FPCs is imposed through constraints on these coefficients. However, the procedure used to estimate the coefficients does not necessarily enforce the appropriate constraints, which was resolved by Peng & Paul (2009). The fundamental assumption of these mixed effects modeling approaches is that the FPCs are in the span of the pre-
specified basis. Inference is sensitive to the choice of basis. For example, if splines are used, the modeling results will be highly impacted by the number of knots and their locations, and estimating these quantities from the data tends to be computationally intractable. If the basis is misspecified, inferences can be invalid, and artifacts of the basis will be present in fitted functions.

Bayesian FPCA methods tend to align with the perspective of [James et al., (2000)] re-lying on the simplicity of representing FPCs with an expansion of a pre-specified basis. Examples include [Behseta et al., (2005), Suarez & Ghosal (2017) and Jiang et al. (2020)], which use Markov chain Monte Carlo (MCMC) algorithms to sample from posterior distributions under alternative Bayesian FPCA models. [van der Linde (2008) and more recently Nolan et al. (2021)] use variational methods to approximate the posterior distribution in this setting. These Bayesian methods have the same drawbacks as the mixed effects modeling approaches, since they are based on the same data generating model. Additionally, the orthogonality constraint of the FPCs is difficult to meet, since it requires prior specification and computational tools for orthogonal functions. Identifiability of the FPCs is either ignored by these approaches or is based entirely on post-hoc processing of MCMC samples or posterior summaries. In contrast, the work of [Jauch et al. (2021)] developed a Bayesian PCA model using the matrix angular central Gaussian distribution to enforce the orthogonality of the PCs. This approach can be applied to FPCA if the functional data are observed on the same dense grid of points, but there is no framework for inference for functions at other points.

In this work we formulate a model for Bayesian FPCA that balances the orthogonality constraint of the FPCs and computational convenience. Our contributions are as follows. (i) We define relaxed mutual orthogonal (ReMO) processes. The joint distribution induced by ReMO processes is governed by a penalty parameter that determines the degree to which the processes are mutually orthogonal. This is achieved by using a single parameter to shrink inferred quantities towards a constrained space, similar in spirit to the Bayesian constraint relaxation framework of [Duan et al. (2019)]. (ii) We demonstrate that ReMO
processes can serve as a foundation for Bayesian FPCA in a wide range of application settings, including sparse and irregular observations and generalized functional data. (iii) We highlight the computational tractability of ReMO processes for FPCA. The specified probability model for individual FPCs is conditionally conjugate for a Gaussian likelihood, which leads to simple MCMC sampling. (iv) FPCA using ReMO processes provides advantages over existing approaches in terms of flexibility, computational convenience and uncertainty quantification. We illustrate that our approach can serve as a layer in a hierarchical model for flexible, interpretable inference with coherent uncertainty quantification.

In Section 2 we define ReMO processes, study their limiting properties, and simulate random samples to gain intuition about their behavior. In Section 3 we specify our model for FPCA using ReMO processes and discuss model extensions to a variety of application settings and computation. We study our methods in comparison with other popular approaches for FPCA using simulated datasets in Section 4, and in Section 5 we illustrate the applicability of our approach using the Cebu Longitudinal Health and Nutrition Survey (Adair et al. 2011). Finally, we conclude with a discussion and directions for future work in Section 6.

2 Relaxed Mutually Orthogonal Processes

With the goal of developing prior models for sets of functions that are consistent with the mutual orthogonality constraint used to define FPCA (Ramsay & Silverman 2005, Ferraty & Vieu 2006), we formulate ReMO processes. For a set of functions, \( \{\lambda_k : T \to \mathbb{R}\}_{k=1}^K \), we seek to enforce a relaxed version of the following constraint:

\[
\langle \lambda_j, \lambda_k \rangle := \int_T \lambda_j(s)\lambda_k(s)ds = 0, \quad j \neq k, \quad (1)
\]

where \( \langle \cdot, \cdot \rangle \) denotes the \( L^2 \) inner product between two functions. Let \( G \) denote the product measure induced by \( K \) independent zero-mean Gaussian processes (GPs) with covariance
kernels $C_k(\cdot, \cdot)$, $k = 1, \ldots, K$, and $\mathcal{H}$ denote the product of reproducing kernel Hilbert spaces defined by the covariance kernels \cite{vanZanten2008}. We refer to these as parent GPs. We assume the measure induced by ReMO processes, denoted by $R$, is absolutely continuous with respect to $G$. We define $R$ through the Radon-Nikodym derivative,

$$
\frac{\partial R}{\partial G}(\lambda_1, \ldots, \lambda_K) \propto \exp\left(-\frac{1}{2\nu_L} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2\right). \tag{2}
$$

The degree to which the orthogonality constraint is met is governed by the penalty parameter, $\nu_L$. The limiting properties of $R$ with respect to $\nu_L$ are summarized by Propositions \ref{prop1} & \ref{prop2}.

**Proposition 2.1.** Let $E := \{\lambda_1, \ldots, \lambda_K \in \mathcal{H} \mid \langle \lambda_j, \lambda_k \rangle \neq 0$, for some $j \neq k\}$. For any measurable subset $E' \subseteq E$, \lim_{\nu_L \to 0} R(E') = 0.

**Proposition 2.2.** For any measurable set $A \subseteq \mathcal{H}$, \lim_{\nu_L \to \infty} R(A) = G(A).

For measurable subsets of non-mutually orthogonal functions, $E'$, $R$ has measure zero, as $\nu_L \to 0$. On the other hand, for any measurable set, $A$, $R$ is equal in measure to $G$, as $\nu_L \to \infty$. Consequently, for fixed $\nu_L > 0$, $R$ balances Gaussian process-like behavior of ReMO processes and mutual orthogonality. As a result, ReMO processes inherit properties from the covariance functions used to define the parent GPs, such as magnitude, smoothness, periodicity, etc. The role of these covariance functions is further explored in Section \ref{sec2.2}.

To study $R$ with fixed $\nu_L > 0$, we define the set of $\omega$-relaxed non-mutually orthogonal functions, $E_\omega := \{\lambda_1, \ldots, \lambda_K \in \mathcal{H} \mid \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 > \omega\}$. For increasing $\omega$, $E_\omega$ will not contain any functions that are close to mutually orthogonal, quantified through the pairwise sum of squared inner products. Proposition \ref{prop3} states the behavior of $R$ for measurable subsets of $E_\omega$. 

5
Proposition 2.3. For any $\nu_\lambda > 0$ and measurable subset $E'_\omega \subseteq E_\omega$ with non-zero $G(E'_\omega)$,

$$\frac{R(E'_\omega)}{G(E'_\omega)} \leq F \exp \left(-\frac{\omega}{2\nu_\lambda} \right),$$

where $F = \left(\mathbb{E}_G[\exp \left(-\frac{1}{2\nu_\lambda} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right)] \right)^{-1}$ is a normalizing constant.

Relative to the Gaussian measure, $R$ assigns probability to $E'_\omega$ that decays exponentially fast as $\omega$ increases, with rate governed by $\frac{1}{2\nu_\lambda}$. Through the definition of $R$ in Equation (1), ReMO processes will tend be closer to mutually orthogonal than Gaussian processes, consistent with our goal of imposing near-mutual orthogonality on a set of functions.

Proofs for Propositions 2.1, 2.2, and 2.3 are presented in Appendix A.

2.1 Conditional Distribution of $\lambda_k$ given $\{\lambda_j\}_{j \neq k}$

A key ingredient of the FPCA model developed in Section 3 is that the functional principal components are assumed a priori to be ReMO processes, $\lambda_1, \ldots, \lambda_K \sim R$. Seemingly, it would appear that inferring the FPCs would be challenging because of the analytically intractable normalizing constant of the measure induced by ReMO processes, $R$, defined in Equation (2). However, the distribution of a single ReMO process $\lambda_k$ given all others $\{\lambda_j\}_{j \neq k}$ has a computationally convenient distribution, stated in Proposition 2.4.

Proposition 2.4. Given $\{\lambda_j\}_{j \neq k}$, $\lambda_k$ is a zero-mean Gaussian process with covariance function

$$C^{\nu_\lambda}_k(s,t) = C_k(s,t) - h_{\Lambda_{(-k)}}(s) \top (\nu_\lambda I_{K-1} + H_{\Lambda_{(-k)}})^{-1} h_{\Lambda_{(-k)}}(t), \text{ where (3)}$$

$$h_{\Lambda_{(-k)}}(t) = \int_T C_k(s,t) A_{(-k)}(s) ds, \quad H_{\Lambda_{(-k)}} = \int_T \int_T C_k(s,s') A_{(-k)}(s) A_{(-k)}(s') \top dsds' \quad \text{(4)}$$

with $A_{(-k)}(s) = (\lambda_1(s), \ldots, \lambda_{k-1}(s), \lambda_{k+1}(s), \ldots, \lambda_K(s)) \top$. 
Proposition 2.4 implies that the finite-dimensional conditional distribution of $\lambda_k$ evaluated at a fixed set of points, given $\{\lambda_j\}_{j \neq k}$, is multivariate Gaussian, which is conditionally conjugate to the Gaussian likelihood of discretely observed functional data assumed in the FPCA model in Section 3. Conditional conjugacy enables posterior inference through a computationally simple Gibbs sampler, outlined in Section 3.4 and detailed further in Appendix B. The behavior of $\lambda_k|\{\lambda_j\}_{j \neq k} \sim GP(0, C^{\nu}_{\lambda_k}(\cdot, \cdot))$ is governed by the covariance function $C^{\nu}_{\lambda_k}(s, t) = \text{cov}(\lambda_k(s), \lambda_k(t)|\{\lambda_j\}_{j \neq k})$, illustrated in Section 2.2.

Conditionally, ReMO processes are related to the orthogonal GPs of Plumlee & Joseph (2018), which were studied in the context of the design and analysis of computer experiments. The goal of their work was to develop a flexible non-parametric model for a single function $f(t)$, which imposes that $f(t)$ is orthogonal to a pre-specified set of functions $g(t) = (g_1(t), \ldots, g_K(t))^\top$, which typically have some physical interpretability. To this end, the authors define orthogonal GPs through an orthogonal covariance function,

$$C^\perp_g(s, t) = C(s, t) - h_g(s)^\top H_g^{-1} h_g(t),$$

where $C(s, t)$ is a fixed covariance function. The authors show that a zero-mean Gaussian process with covariance function $C^\perp_g(\cdot, \cdot)$, $f \sim GP(0, C^\perp_g(\cdot, \cdot))$, is orthogonal to each $g_1(t), \ldots, g_K(t)$ with probability 1. Within the context of our model, for a given $k$, we consider $g(t) = \Lambda(-k)(t)$ and the orthogonal Gaussian process $\lambda_k|\{\lambda_j\}_{j \neq k} \sim GP(0, C^\perp_{\Lambda(-k)}).$

The relationship between ReMO processes and orthogonal GPs is stated in Proposition 2.5.

**Proposition 2.5.** $\lambda_k|\{\lambda_j\}_{j \neq k} \overset{d}{\to} \lambda^\perp_k|\{\lambda_j\}_{j \neq k}$ as $\nu_\lambda \to 0$.

We emphasize the fundamental difference between ReMO processes and orthogonal GPs is that an orthogonal GP is a prior for a single function restricted to be orthogonal to a fixed set of functions, whereas ReMO processes provide a prior for a collection of near-mutually orthogonal functions. When studying orthogonal GPs, the authors in Plumlee & Joseph...
Figure 1: (a) The value of $\lambda_1$ on which $\lambda_2$ is conditioned. Boxplots of 1000 realizations of $\langle \lambda_1, \lambda_2 \rangle$, with $\lambda_2 | \lambda_1 \sim GP(0, C_2^{\nu_\lambda}(\cdot, \cdot))$ when (b) $\tau_2^2 = 1$, $\nu_\lambda = 0.0001$ are fixed as $l_2 = 0.001, 0.01, 0.1, 1$ varies, (c) $l_2 = 0.01$, $\nu_\lambda = 0.0001$ are fixed as $\tau_2^2 = 0.25, 0.5, 1, 2$ varies, (d) $l_2 = 0.01$, $\tau_2^2 = 1$ are fixed as $\nu_\lambda = 0.0001, 0.01, 1, 100$ varies.

Figure 2: Realizations of $\lambda_2 | \lambda_1 \sim GP(0, C_2^{\nu_\lambda}(\cdot, \cdot))$ with $\tau_2^2 = 1$, $\nu_\lambda = 0.0001$ fixed, while (a) $l_2 = 0.001$, (b) $l_2 = 0.01$, (c) $l_2 = 0.1$, (d) $l_2 = 1$.

(2018) restrict attention to orthogonal Gaussian processes with continuous sample paths. Proposition 2.6 follows under an analogous assumption that the parent GPs used to define ReMO processes have continuous sample paths, which verifies the desired property that for any $j \neq k$ the inner product between $\lambda_j$ and $\lambda_k$ converges in probability to zero.

**Proposition 2.6.** $\langle \lambda_j, \lambda_k \rangle | \{\lambda_j\}_{j \neq k} \xrightarrow{P} 0$ for any $j \neq k$ as $\nu_\lambda \to 0$

Proofs for Propositions 2.4, 2.5, and 2.6 are presented in Appendix A.

### 2.2 Role of Covariance Parameters

While the joint distribution defined in Equation (2) is well defined for any choice of covariance kernels $C_k(\cdot, \cdot)$, $k = 1, \ldots, K$, we choose to illustrate the model based on the squared
exponential covariance kernel (Rasmussen & Williams 2006), defined as

$$C_k(s, t) = \tau_k^2 \exp \left\{ -\frac{1}{2l_k^2} (s - t)^2 \right\}, \quad s, t \in T \times T.$$  \hspace{1cm} (5)

In a typical Gaussian process setting, the scale parameters $\tau_k^2$ determine the spread of realized processes about the mean, while the length-scale parameters $l_k$ determine the smoothness of realizations.

For a simple setting when $K = 2$, we illustrate the conditional prior of Proposition 2.4 when $\lambda_1$ is sampled from a Gaussian process, shown in panel (a) of Figure 1 and $C_2$ is defined according to Equation (5). Panels (b)-(d) shows the sensitivity of the inner product between realizations of $\lambda_2|\lambda_1$ and $\lambda_1$ under different parameter settings, which are further detailed in the following paragraphs. The conditional variance, $C^{\nu_\lambda}_2(\cdot, \cdot)$, is governed by 3 parameters, $l_2, \tau_2^2$, and $\nu_\lambda$. To gain insight into the role of these parameters, we sample
realizations $\lambda_2|\lambda_1 \sim GP(0,C_2^{\nu}(\cdot,\cdot))$ when two of the parameters are held constant while the third varies.

Figure 2 shows realizations from the prior with $\tau^2 = 1, \nu_\lambda = 0.0001$ fixed, while $l^2_2 = 0.001, 0.01, 0.1, 1$ varies in panels (a)-(d). Notice, as the length-scale parameter increases, the realizations become more smooth. When the realizations are relatively smooth, in panels (c)&(d), the variance shrinks around $t \approx .65$. This feature is present because of the prior’s need to offset the extrema of $\lambda_1$ to enforce orthogonality between the two functions. When the realizations are relatively rough, the shrinking of the variance of $\lambda_2$ is less evident, since rougher functions tend to be close to orthogonal than smoother ones. Panel (b) of Figure 1 shows a boxplot of the inner product between $\lambda_1$ and 1000 realizations of $\lambda_2$ for the different values of the length-scale parameter. Varying $l^2_2$ does not appear to have an effect on the inner product between $\lambda_1$ and the realizations of $\lambda_2|\lambda_1$.

Figure 3 shows realizations from the prior with $l^2_2 = .1, \nu_\lambda = 0.0001$, and $\tau^2_2 = 0.25, 0.5, 1, 2$ varying in panels (a)-(d). Varying this parameter changes the spread of the realizations, more drastically outside of the region where $t \approx .65$. Panel (c) of Figure 1 shows a boxplot of the inner product between $\lambda_1$ and 1000 realizations of $\lambda_2|\lambda_1$ for the different values of the scale parameter. Again, varying $\tau^2_2$ does not appear to have an effect on the inner product between the $\lambda_1$ and the realizations of $\lambda_2|\lambda_1$.

Figure 4 shows realizations from the prior with $l^2_2 = .1, \tau^2_2 = 1$ fixed, and $\nu_\lambda = 0.0001, 0.01, 1, 100$ varying in panels (a)-(d). As previously discussed in Section 2.1 this parameter has an important role in enforcing the orthogonality constraint. As the parameter increases, in panel (d), the realizations have typical Gaussian process behavior. However, as the parameter decreases, the realizations become close to orthogonal to $\lambda_1$. Panel (d) of Figure 1 shows a boxplot of the inner product between $\lambda_1$ and 1000 realizations of $\lambda_2|\lambda_1$ for the different values of the penalty parameter. As expected, varying $\nu_\lambda$ has a drastic effect on the inner product between $\lambda_1$ and the realizations of $\lambda_2|\lambda_1$. The inner product is negligible when $\nu = 0.0001$ compared to when $\nu = 100$.

The parameters $l^2_2$ and $\tau^2_2$ retain their meaning from the typical Gaussian process setting.
of governing the smoothness and spread of the prior, respectively. The penalty parameter $\nu_\lambda$ only has a role in enforcing the orthogonality constraint. Throughout the implementations of our model implemented in this work, we infer the parameters $l_k^2$ and $\tau_k^2$, while fixing $\nu_\lambda$ to be a small value.

3 FPCA using Relaxed Mutually Orthogonal Processes

To fully specify a Bayesian model for FPCA, we begin by stating an observation model in Subsection 3.1 followed by prior formulation of model parameters in Subsection 3.2. We extend our model to generalized FPCA (GFPCA) in Subsection 3.3.

3.1 Observation Model

Let $y_i : T \rightarrow \mathbb{R}, i = 1, \ldots, n$ denote functional observations. We specify an observation model that is consistent with a typical FPCA approach (Ramsay & Silverman 2005),

$$y_i(t) = \mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i + \epsilon_i(t), \quad i = 1, \ldots, n.$$  (6)

In Equation (6), $\mu(t)$ represents a mean process common to all observations, the FPC basis $\lambda_1(t), \ldots, \lambda_K(t)$ are latent functions, $\eta_i \in \mathbb{R}^K$ are the latent FPC scores, and $\epsilon_i(t)$ denotes a subject specific error term. Mutual orthogonality is imposed on $\lambda_1(t), \ldots, \lambda_K(t)$ to ensure identifiability of the FPCs, and the FPC scores, $\eta_i \in \mathbb{R}^K$, are constrained to have mean zero across observations for identifiability of the mean process. The novel prior models that we use to meet these constraints are detailed in Section 3.2.

In the setting of sparse and irregular observations, it is assumed that each $y_i$ is observed on an observation-specific grid $t_i = (t_{i,1}, \ldots, t_{i,m_i})^\top$. The FPCA model of Equation (6) can
be altered to accommodate sparse and irregular observations,

\[ y_i(t_i) = O_i \mu(t) + O_i(\lambda_1(t), \ldots, \lambda_K(t))^{\top} \eta_i + \epsilon_i(t_i), \quad i = 1, \ldots, n. \quad (7) \]

We use \( t \) to denote a common grid of size \( m \) computed using the collection of all observation-specific grid points. In general, we use \( f(t) = (f(t_1), \ldots, f(t_m))^{\top} \) to denote a vector of function evaluations. The matrix \( O_i \in \{0, 1\}^{m_i \times m} \) relates the common and observation-specific grids, so that \( O_i f(t) = f(t_i) \). The \( j \)th row of \( O_i \) is a row vector of 0’s with a 1 in the index where \( t \) is equal to \( t_{i,j} \).

For our Bayesian FPCA model, we assume a Gaussian white noise error process, so that \( \epsilon_i \overset{iid}{\sim} GP(0, C_\epsilon(\cdot, \cdot)) \). For \( s, t \in T \), \( C_\epsilon(s, t) = \sigma^2 \delta(s - t) \), where \( \delta(s - t) \) is the Dirac delta function. For the discretized version of the model presented in Equation (7), the assumed error process induces a Gaussian likelihood,

\[ y_i | \mu(t), \lambda_1(t), \ldots, \lambda_K(t), \eta_i \overset{ind}{\sim} N_{m_i}(O_i \mu(t) + O_i(\lambda_1(t), \ldots, \lambda_K(t))^{\top} \eta_i, \sigma^2 I_{m_i}). \]

### 3.2 Prior Specification

We assume a zero-mean Gaussian process \textit{a priori} for the mean term in Equation (6), \( \mu \overset{\sim}{\sim} GP(0, C_\mu(\cdot, \cdot)) \). This is a flexible and computationally convenient prior, since the full conditional distribution for the discretized process, \( \mu(t) \), can be sampled from directly. We assume a squared exponential covariance kernel for \( C_\mu(\cdot, \cdot) \) that is governed by a smoothness parameter \( l_\mu \) and scale parameter \( \tau_{\mu}^2 \). For the error variance, we assume a conditionally conjugate prior \( \sigma^2 \sim \text{inverse-gamma}(\alpha_\sigma, \beta_\sigma) \).

We assume the latent FPCs of equation (6) are ReMO processes, \( \lambda_1, \ldots, \lambda_K \sim R \). We set \( \nu_\lambda = 1 \times 10^{-4} \) when implementing our model in Sections 4 & 5. We provide a detailed sensitivity analysis of \( \nu_\lambda \) for a simulated dataset in Appendix C. In general, \( \nu_\lambda \) should be chosen to be small enough so that the FPCs are able to be identified. However, choosing
\( \nu \) extremely small can lead to slow MCMC mixing. The choice of \( \nu = 1 \times 10^{-4} \) reflects a reasonable balance between these two extremes in our experience.

FPC Scores are typically forced to have zero mean \cite{RamsaySilverman2005, FerratyVieu2006} so that the mean process is identifiable. Otherwise, the FPCs and the mean could be conflated. Let \( \eta_k = (\eta_{1,k}, \ldots, \eta_{n,k})^\top \) denote the vector of the \( k \)th FPC scores for all observations. In order to identify the mean process in Equation (6), we impose a relaxed version of the sum-to-zero constraint,

\[
\sum_{i=1}^{n} \eta_{i,k} \mathbf{1}_n = \eta_k^\top \mathbf{1}_n = 0, \text{ for all } k = 1, \ldots, K.
\]

This is accomplished via a prior that favors \( \eta_k \) values with small sums,

\[
\pi(\eta_k) \propto \exp\left(-\frac{1}{2} \eta_k^\top \eta_k \right) \exp\left\{-\frac{1}{2\nu} \left(\mathbf{1}_n^\top \eta_k\right)^2\right\}, \tag{8}
\]

independently across \( k \). The first term in the prior corresponds to the kernel of a multivariate Gaussian distribution with identity covariance matrix. The second term is used to enforce the (relaxed) sum-to-zero constraint. When \( \nu \) is very small, the prior puts little weight on \( \eta_k \) vectors that have non-zero sum and favors \( \eta_k \) vectors that approximately sum to zero.

Through rearranging the terms in (8), the prior can be recognized as a zero-mean multivariate Gaussian distribution with covariance,

\[
\text{var}(\eta_k) = \left(I_n + \frac{1}{\nu} \mathbf{1}_n \mathbf{1}_n^\top\right)^{-1} = I_n - \frac{1}{\nu} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{\nu + n}.
\]

The matrix inversion is computed via Woodbury’s Formula \cite{Harville1998}. The limiting distribution of \( \eta_k \), as \( \nu \to 0 \), is stated in Proposition 3.1

**Proposition 3.1.** \( \eta_k \overset{d}{\to} (I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top)N(0, I_n) \) as \( \nu \to 0 \).
Note that \((I_n - \frac{1}{n}1_n1_n^\top)\) is the projection matrix that projects \(n\)-vectors onto the subspace of vectors that sum to zero. Consequently, when \(\nu_n\) is small, the prior effectively enforces the sum-to-zero constraint. A proof for this Proposition is presented in Appendix A.

3.3 Generalized Functional Principal component Analysis

GFPCA extends FPCA to allow non-Gaussian (e.g., binary or count) observations. Typically, the expected values of the data are modeled through a known link to a real-valued function that is represented through an FPC basis. Hall et al. (2008) extends Yao et al. (2005) for GFPCA by using a latent Gaussian process. Alternatively, there are both Bayesian (van der Linde 2009) and frequentist (Goldsmith et al. 2015, Wrobel et al. 2019, Zhong et al. 2021) approaches in which elements of the FPC basis for the latent real-valued function are an expansion of a pre-specified basis. ReMO processes avoid the need to pre-specify such a basis, and we outline the corresponding GFPCA details below.

To modify (6), to allow non-Gaussian \(y_i(t)\), let

\[
E[y_i(t)|\mu, \lambda_1, \ldots, \lambda_K, \eta_i] = h\{\mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i\}, \quad i = 1, \ldots, n, \tag{9}
\]

where \(h(\cdot)\) is an appropriate monotone one-to-one and differentiable link function. Model (9) implies the following model for irregular observed data,

\[
E[y_i(t_i)|\mu, \lambda_1, \ldots, \lambda_K, \eta_i] = h\{O_i\mu(t) + O_i(\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i\}, \quad i = 1, \ldots, n.
\]

In the application in Section 5, we consider binary functional data with \(h(\cdot)\) the standard normal CDF, \(\Phi\).
3.4 Notes on Computation

The target posterior distribution for FPCA is

\[
\pi(\mu(t), l_\mu, \tau_\mu^2, \{\lambda_k(t), l_k, \tau_k^2, \eta_k, \psi_k\}_{k=1}^K, \sigma^2|\{y_i(t_i)\}_{i=1}^n) \propto \\
\ell(\{y_i(t_i)\}_{i=1}^n|\mu(t), l_\mu, \tau_\mu^2, \{\lambda_k(t), l_k, \tau_k^2, \eta_k, \psi_k\}_{k=1}^K, \sigma^2) \times \\
\prod_{k=1}^K \pi(\lambda_k|l_k, \tau_k^2) \pi(l_k, \tau_k^2) \pi(\eta_k|\psi_k) \pi(\psi_k) \times \\
\pi(\mu|l_\mu, \tau_\mu^2) \pi(l_\mu, \tau_\mu^2) \pi(\sigma^2).
\]

We concisely restate the likelihood and prior distributions in hierarchical model notation in Appendix D. The \(\psi_k\) auxiliary parameters introduced in the priors for \(\eta_k\) govern prior variance following the parameter expansion approach of Ghosh & Dunson (2009), which introduced this technique in the context of multivariate latent factor models. Inference of the \(\psi_k\) parameters themselves is usually not of interest. Saved posterior samples of \(\psi_k\) are used to scale saved posterior samples of \(\lambda_k\) and \(\eta_k\). This approach has been shown to diminish posterior dependence and lead to better MCMC mixing.

We present an adaptive Metropolis-within-Gibbs algorithm to draw \(N\) many samples from the joint posterior in Algorithm 1 in Appendix C. Gibbs steps are available for the majority of parameters, and we use adaptive Metropolis steps with proposal variances that are tuned to meet targeted acceptance rate (Andrieu & Thoms 2008, Algorithm 4) for parameters that do not have conditionally conjugate full posterior distributions. In Algorithm 2 of Appendix C, we present an extension of Algorithm 1 for the case of sparse and irregularly sampled binary functional data.

When evaluating the conditional variance \(C_{\nu,\lambda_k}^{\nu,\lambda}(\cdot,\cdot)\) in Equation (14), we approximate integrals in Equation 15 using Riemann sums. Specifically, we use the approximation \(\int_T f(s)ds \approx \sum_{l=1}^m w_l f(t_l)\) relying on the grid \(t = (t_1, \ldots, t_m)^T\). The weights \(w_l, l = 1, \ldots, m\), are scalars accounting for the spacing of the grid. Letting \(W = \text{diag}(w_1, \ldots, w_m)\), we choose \(w_1 = \frac{t_2-t_1}{2}, w_l = \frac{t_{l+1}-t_{l-1}}{2}, l = 2, 3, \ldots, m-1, w_m = \frac{t_m-t_{m-1}}{2}\). The approximation
error, \(| \int_T f(s)ds - \sum_{i=1}^m w_i f(t_i) |\), is bounded above by a value proportional to \( \frac{1}{m^2} \).

The full conditional distribution of \( \lambda_k(t) \) given all other parameters in the model and the data is given by

\[
\lambda_k(t) \mid - \sim N(\mathbb{E}[\lambda_k(t)\mid -], \mathbb{V}[\lambda_k(t)\mid -]),
\]
\[
\mathbb{V}[\lambda_k(t)\mid -] = \left\{ C_k^{-1}(t, t) + \frac{1}{\nu_k} W \Lambda_k(-k)(t) \Lambda_k(-k)(t)^T W + \frac{1}{\sigma^2} \sum_{i=1}^n \eta_{i,k}^2 O_i^T O_i \right\}^{-1}
\]
\[
\mathbb{E}[\lambda_k(t)\mid -] = \frac{1}{\sigma^2} \mathbb{V}[\lambda_k(t)\mid -] \sum_{i=1}^n \left[ \eta_{i,k} O_i^T \left\{ y_i(t_i) - O_i \mu(t) - O_i \sum_{j \neq k} \eta_{i,j} \lambda_j(t) \right\} \right].
\]

This Gibbs step used to sample the FPCs is simple, highlighting a large appeal for using ReMO processes for FPCA: MCMC samples from full conditional distributions are easily generated. This step is easily modified to GFPCA with binary or categorical observations by using data augmentation [Albert & Chib (1993)].

In the models formulated thus far, there is non-identifiability of the order and signs of the FPCs, which can result in misalignment of saved MCMC samples. To resolve this, we use the algorithm for label switching and sign ambiguity from [Poworoznek et al. (2021)].

Code to implement the FPCA and GFPCA models on simulated data is available on GitHub: [https://github.com/jamesmatuk/ReMO-FPCA](https://github.com/jamesmatuk/ReMO-FPCA). This repository also contains implementations of all examples presented in this paper.

### 4 Simulation Experiments

In this section, we use simulation experiments to study the efficacy of ReMO processes as the foundation for FPCA and GFPCA by studying estimation performance as well as the frequentist coverage of regression coefficients within the context of latent variable regression compared to competing methods.

In Appendix E, we present an additional simulation, where data are generated from an FPCA model inferred from the Berkeley growth dataset [Ramsay & Silverman (2005)].
assumed generative FPCA model coincides with that of Yao et al. (2005). Even in this realistic simulation setting under model misspecification, our FPCA model using ReMO processes performs well compared to competing methods.

4.1 Function Estimation in FPCA

To study ReMO processes for FPCA, we generate $n = 100$ functions on a grid, $t$, with $m = 30$ equally spaced points, according to

$$f_i(t) = \mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i, \ i = 1, \ldots, n.$$
With $K = 2$, we generate $\mu$, $\lambda_1$, $\lambda_2$, $\eta_1$, $\eta_2$ according to the priors presented in Section 3.2 with $l_\mu^2 = l_{\lambda_1}^2 = l_{\lambda_2}^2 = .4$ and $\tau_\mu^2 = \tau_{\lambda_1}^2 = \tau_{\lambda_2}^2 = 1$. We study the effect of sparsity level on the estimation results by randomly omitting 25%, 50%, and 75% of the observations, so that the simulated noisy functions are observed on $t_i$, $i = 1 \ldots, n$. We measure estimation performance using mean integrated squared error $MISE = \frac{1}{n} \sum_{i=1}^{n} \int_{T} (\hat{f}_i(s) - f_i(s)) ds$, where $\hat{f}_i$ is an estimate of $f_i$.

We use the posterior mean of $\mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i$ to produce an estimate of $f_i$ based on 1,000 MCMC samples. In terms of other FPCA methods, we use the R package `fdapace` to estimate $f_i$ according to the principal analysis through conditional expectation (PACE) methodology (Yao et al. 2005). PACE uses kernel smoothing to estimate a mean and covariance function for sparse functional data, from which FPCA is carried out. Additionally, we use the R package `fpca` to implement the methodology of Peng & Paul (2009), termed ‘Newton’ in their paper. The Newton methodology uses a pre-specified basis to represent FPCs, and inference is carried out through the basis coefficients. The work of Crainiceanu & Goldsmith (2010) provides implementation of a Bayesian approach to FPCA. In an empirical Bayes setup, the authors mean-center the observations, and estimate a covariance surface using penalized thin-plate splines, from which estimates of FPCs are computed. While fixing the FPCs to their estimated values, the authors formulate priors for FPC scores and error variance, and perform MCMC using WinBUGS. We term this method ‘CG’.

In all methods for FPCA, we misspecify $K = 5$, and we replicate the entire data generating mechanism 100 times for comparison. Figure 5 shows average MISE results for $\sigma^2 = .25$ and $\sigma^2 = 1$ for different levels of sparsity in panels (a)&(b), respectively. For all of the methods, the estimation performance deteriorates as the sparsity level and noise increases, since there is less information to estimate the underlying functions. Consistently, the estimation error from ReMO is on average lower than the others. However, PACE and ReMO are very close in the high noise setting. We note that the methods Newton and CG that base their estimates of FPCs on pre-specified bases generally have significantly worse
performance.

4.2 Function Estimation in GFPCA

To study ReMO processes in the GFPCA setting, we generate $n = 100$ binary functions on a grid $t$ with $m = 30$ equally spaced points according to

$$P(y_i(t) = 1) = \Phi(f_i(t)) = \Phi(\mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i), \ i = 1, \ldots, n.$$

With $K = 2$, we generate $\mu$, $\lambda_1$, $\lambda_2$, $\eta_1$, $\eta_2$ according to the priors presented in Section 3.2 with $l^2_\mu = l^2_{\lambda_1} = l^2_{\lambda_2} = .4$. Under different settings in the simulation, we study when the scale of the mean and FPCs are relatively small $\tau^2_\mu = \tau^2_{\lambda_1} = \tau^2_{\lambda_2} = 1$ or relatively large $\tau^2_\mu = \tau^2_{\lambda_1} = \tau^2_{\lambda_2} = 10$. In the larger scale setting, $P(y_i(t) = 1)$ is likely to be much closer to 0 or 1 compared to the smaller scale setting. Again, to study the effect of sparsity level on the estimation results, we randomly omit 25%, 50%, and 75% of the observations, so that the binary observations are observed on $t_i, \ i = 1 \ldots, n$. Estimation performance of $\phi(f_i), \ i = 1, \ldots, n$ is measured through MISE.

From our model, we use the posterior mean of $\phi(\mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i)$, to produce an estimate of $\phi(f_i)$ based on 1,000 MCMC iterations. In terms of other methods for comparison, we use the bfpca method from the registr package in R to estimate $f_i$ using the methodology described in Wrobel et al. (2019). We also estimate $f_i$ using the sparse logistic FPCA (SLFPCA) methods of Zhong et al. (2021) from the SLFPCA R package. Both of these methods specify the FPCs as an expansion of a pre-specified basis, and inference is based on the coefficients. We misspecify $K = 5$ for all of the GFPCA methods, and we replicate the entire data generating mechanism 100 times for comparison. Panels (c) & (d) of Figure 5 shows average MISE results for the lower and higher scale settings, respectively. The estimation errors tend to increase with sparsity level. The MISE from ReMO is consistently below that of the others.
4.3 Frequentist Coverage of Regression Coefficients

Finally, we study our model’s ability to serve as a foundation for inferential tasks in a latent factor regression setting (Montagna et al. 2012). Specifically, data are generated as in Section 4.1 with \( K = 1 \) and

\[ \eta_{i,1} = \Theta x_{i,1} + \xi_{i,1}, \]

where \( x_{i,1} \) is a fixed covariate, \( \Theta \) is a regression coefficient and \( \xi_{i,1} \sim \text{iid} N(0,1) \). In each of 100 replicated simulations, \( \Theta \) and \( x_{i,1} \) are independently generated as standard normal random variables. We view \( \Theta \) as a parameter on which to do inference while \( x_{i,1}, i = 1, \ldots, n \), are fixed. We study frequentist coverage of a 95\% credible interval for \( \Theta \) based on 10,000 MCMC samples for our model. For comparison, we estimate \( \eta_{i,1}, i = 1, \ldots, n \), using PACE and Newton from Section 4.1 and regress the estimated FPC scores onto the covariate. For CG, the FPC score regression is a hierarchical layer in the model. Based on the regression inference, the coverage of 95\% confidence intervals can be studied. In all models, we misspecify \( K = 2 \) and replicate the entire data generating mechanism 100 times to study coverage. Figure 5 panels (e) & (f) show the coverage based on the replicates for different noise levels \( \sigma^2 = .25 \) and \( \sigma^2 = 1 \), respectively. In the low noise setting, all methods provide reasonable coverage, except for PACE in the high sparsity setting. The high noise setting illustrates the importance of uncertainty propagation between estimation and inference. In Newton and PACE, the coverage falls extremely low because of overconfidence due to conditioning on the estimation results. The low coverage from CG is due to the empirical Bayes setup, which fails to account for uncertainty in the FPCs. In either setting ReMO propagates uncertainty between estimation and inference improving performance.

5 Cebu Longitudinal Health and Nutrition Survey

The Cebu Longitudinal Health and Nutrition Survey is an ongoing multi-generational study (Adair et al. 2011). The original purpose was to study maternal, child, and environmental
Figure 6: (a) Weight by age for the $n = 2898$ children in the Cebu Longitudinal Health and Nutrition Survey. (b) Proportion of missing weight data by age.

Factors and their connection to health, social, and economic outcomes for mother-child pairs in the metropolitan area of Cebu, Philippines. The sample has expanded through the years to include the offspring of the children in the original study. The dataset and background information are available at [https://cebu.cpc.unc.edu/](https://cebu.cpc.unc.edu/).

We focus on a subset of the data that corresponds with the original mother-child pairs. The weight of the infants as well as other covariates were recorded in 2 month intervals from birth to the time the child was 2 years old. Previous work [Adair et al. 1993, Adair & Guilkey 1997] found that scalar covariates, such as height of the mother, sex of the child, area (urban or rural) where the family lives, season of birth (rainy or dry), and longitudinal covariates, such as breastfeeding status and illness, are associated with the weight dynamics in early childhood. We develop a flexible model for weight dynamics, while accounting for the scalar and longitudinal covariates as follows using our methodology for FPCA and GFPCA,

$$y_i(t_i) = O_i\mu(t) + O_i(\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i + O_i \sum_{j=1}^4 \int_T \beta_j(s, t) z_{i,j}(s) ds + \epsilon_i(t_i)$$

$$\eta_i = \Theta x_i + \xi_i$$

The longitudinal recordings of weight of the $i^{th}$ child are denoted by $y_i(t_i)$, where the
observation-specific grid points $t_i$ are a subset of $t = (0, \frac{1}{2}, \ldots, \frac{24}{2})^\top$, representing the two month intervals on which the measurements are recorded. We use our FPCA approach developed in Section 3 for the priors on $\mu$, $\lambda_1, \ldots, \lambda_K$ and $\xi_i$.

The response can vary with the scalar covariates, $x_i \in \mathbb{R}^4$, mentioned in the previous paragraph by regressing the FPC scores in a latent factor regression (Montagna et al. 2012). The coefficient matrix $\Theta \in \mathbb{R}^{K \times 4}$ models the relationship between the FPC scores and the scalar covariates. To account for longitudinal covariates, $z_{i,j}(s)$, $j = 1, \ldots, 4$, we use a functional linear model (Ramsay & Dalzell 1991). These covariates are sparsely recorded and track if a child is fed breast milk and 3 indicators of illness (diarrhea, fever, and cough).

We use our GFPCA approach for model-based imputation of these covariates for the functional linear model to be well defined. For the coefficient surface $\beta_1(s, t)$ corresponding to the breast milk indicator, we specify $\beta_1(s, t) = 0$ for $s > t$ in a historical linear model approach (Malfait & Ramsay 2003). Through this setup, weight at time $t$ is only affected by breastfeeding status at earlier ages. For the coefficient surface of the three illness indicators $\beta_2(s, t)$, $\beta_3(s, t)$, $\beta_4(s, t)$, we specify $\beta_j(s, t) = 0$ for $s \neq t$, $j = 2, 3, 4$ in a concurrent linear model (Hastie & Tibshirani 1993). Through this setup the illness status of a child at time $t$ only affects the weight at time $t$. We fit the coefficient surfaces using sensible pre-specified bases. Full details of the hierarchical model specification are presented in Appendix F.1.

We base inference for model parameters on $n = 2898$ mother-child pairs, which is a large portion of the 3327 pairs originally enrolled in the study. We excluded subjects due to child death, the family moving out of the Cebu metropolitan area, refusal to participate after the baseline survey, or if there was absolutely no covariate or response data recorded. In the remaining $n = 2898$ subjects, the sparsity in the response and longitudinal covariates tends to arise due to a missed survey or failure to answer a specific question. Posterior inference is based on 100,000 MCMC samples (to be conservative), where convergence is assessed using trace plot diagnostics, shown in Appendix F.5.

Figure 6 panel (a) displays the weight of all children measured every two months from birth to age 2 and panel (b) displays the proportion of missing observations by age. Pos-
Figure 7: (a) Posterior samples of the mean process, $\mu(t)$, and (b)-(f) FPCs, $\lambda_1(t), \ldots, \lambda_5(t)$, for weight measurements of children recorded in the Cebu Longitudinal Health and Nutrition Survey.

Posterior samples of the mean process, $\mu$, and FPCs, $\lambda_1, \ldots, \lambda_5$ are displayed in Figure 7 in panels (a) and (b)-(f), respectively. We selected $K = 5$ based on visual inspection of scree plots, shown in Appendix F.2. The mean process increases rapidly from age 0 to 6 months, then appears to behave nearly linearly from age 6 months to 2 years. The first FPC shown in panel (b) captures subject specific deviations having roughly a similar shape as the mean. The latter FPCs in panels (c)-(f) capture more intricate variability that is smaller in magnitude.

The effect of the scalar covariates on the response is modeled through the FPC scores $\eta_i = \Theta x_i + \xi_i$. With all other variables held constant, the effect of a unit increase of a single covariate on the response is given by $(\lambda_1(t), \ldots, \lambda_5(t))^\top \Theta_q$, where $\Theta_q$ denotes the $q^{th}$ column of the matrix $\Theta$, for $q = 1, \ldots, 4$. Posterior samples of these functions are shown in Figure 8. Panel (a) shows the inferred effect of a unit increase in mother’s height on child weight. The model infers that children with taller mothers generally weigh more.
Figure 8: (a)-(d) Posterior samples of $(\lambda_1(t), \ldots, \lambda_5(t))^{\top} \Theta_q$ for $q = 1, \ldots, 4$, representing the effect of the scalar covariates (height of the mother, sex of the child, area (urban or rural) where the family lives, season of birth (rainy or dry)) on weight. Mother’s height and sex of child are clearly associated with weight dynamics. Strata and season of birth appear to have near zero effects with relatively high uncertainty.

Panel (b) shows the inferred difference in weight between male and female children, where $x_{i,2} = 0$ if the $i$th subject is male. The model infers that female children generally weigh less than male children. The discrepancy in weight is small at birth, but changes rapidly from birth to 6 months. Panel (c) shows the inferred difference between children born in urban or rural strata, where $x_{i,3} = 0$ if the $i$th subject was born in an urban area. Panel (d) shows the inferred difference between children born in the rainy or dry season, where $x_{i,4} = 0$ if the $i$th subject was born during the rainy season. Compared to mother’s height or sex, these covariate effects are smaller in magnitude with greater uncertainty.

In order to compute the integral $\int_T \beta_1(s,t) z_{i,1}(s) \, ds$ related to the functional linear model relating breastfeeding status to weight, we rely on GFPCA for model based imputation of the binary functional data $z_{i,1}, i = 1, \ldots, n$. The specifics of the model based imputation are detailed in Appendix F.1. Figure 10 panels (a)-(c) visualize the variability in breastfeeding dynamics captured through our GFPCA model by showing posterior samples of plus and minus one standard deviation in the direction of the FPCs interpreted on the probability scale through the link function $\Phi(\cdot)$. The FPCs mostly capture if a child was breastfed at all and if so, when breastfeeding was discontinued. There are some atypical patterns of breastfeeding trajectories, where children are reported to have been breastfed at a young age, stopped breastfeeding, then continued breastfeeding at a later age. These outliers are further discussed in Appendix F.2. In the posterior samples of the
mean, the values are lower than at 2 months. This is due to the fact that breastfeeding status at birth was measured within the first 2 days of birth, and it could take longer than 2 days for mothers to express breast milk for the first time after childbirth (Casey et al. 1986). Panels (d)-(f) of Figure 9 shows posterior samples of underlying processes describing the probability that a child is fed breast milk by age for different subjects. In panel (d), breastfeeding status was only recorded at 4 ages (black dots). Between the ages when the child was reported to have breast milk, the model infers that the child was fed breast milk. After the last recording, the model is uncertain if and when the child stopped. In panel (e), initially the child was reported to not have breast milk, so the model infers that the child was not given breast milk after the age of the last recording. In panel (f), the child was reported to have breastfed early on, and stopped at some age after age 1. The model is uncertain when breastfeeding was discontinued.
Figure 10: (a)-(d) Posterior samples of $\int_{T} \beta_1 1_{s \leq s'} ds$ for $s' = 0, 0.5, 1, 2$ years, representing the effect of breastfeeding for different durations on weight. (e) Boxplots of posterior samples of $\| \int_{T} \beta_1 1_{s \leq s'} ds \|_2$ for different values of $s'$, representing the magnitude of the effect of breastfeeding for different durations. Children that are fed breast milk tend to weigh less than non-breast milk fed counterparts. The effect of breastfeeding on weight dynamics tends to level off after 1 year.

In order to understand the effects of breastfeeding on child weight, we study posterior samples of the functions $\int_{T} \beta_1 1_{s \leq s'} ds$, where $1_{s \leq s'}$ denotes the indicator function. These functions capture the effect of breastfeeding on weight at $s'$ years with other variables held constant. They can be compared to the effect of never breastfeeding captured by the zero function. Figure 10 shows the effect of breastfeeding for $s' = 0, 0.5, 1, 2$ years. If a child is breastfed only at birth, the effect is close to a zero effect (panel (a)). While breastfed and non-breastfed children are similar in weight at a younger age, their discrepancy is pronounced at larger ages, where breastfed children tend to weigh less. This difference is exaggerated the longer that a child is breastfed (panels (b) through (d)). The effect of breastfeeding tends to level off, as shown through $\| \int_{T} \beta_1 1_{s \leq s'} ds \|_2$ for different values of $s'$ shown in panel (e). Continuing breastfeeding for more months in the first year has a significant impact on weight, while continuing longer after age 1 has a relatively small impact. These inferences corroborate past work devoted to studying the effects of breastfeeding. The work of [Dewey] (1998) is a meta-analysis that studies the effects of breastfeeding on various child growth outcomes, including weight, height and head circumference. In general, they report that breastfed infants tend to self-regulate food intake, have lower observable metabolic rates, and are leaner than their non-breastfed counterparts. The differences in weight between breastfed and non-breastfed children tend to be noticeable after 6 months of age.
Figure 11: (a)-(c) Posterior samples of $\Phi(\mu_{zi}(t))$ for $j = 2, 3, 4$, representing the proportion of children in the study that experienced an illness by age. (d)-(f) Posterior samples of $\beta_j(t)$, $j = 2, 3, 4$, representing the effect of experiencing an illness on weight. Children tend to experience cough more often than fever and fever more often than diarrhea. Experiencing diarrhea or fever tend to have a larger effect on weight than experiencing cough.

For the illness longitudinal covariates we assume a concurrent regression model, which simplifies the integrals, $\int T \beta_j(s, t) z_{ij}(s) ds = \beta_j(t) z_{ij}(t)$, for all $j = 2, 3, 4$, $i = 1 \ldots, n$. Similar to the breastfeeding functions, the illness indicators are sparsely recorded so we rely on model based imputation in this functional linear model term. We outline details in Appendix F.1. In contrast, there does not appear to be structured variability in timing of illness, with children experiencing diarrhea, fever and cough sporadically. Consequently, model based imputation is based only on a mean process interpreted on the probability scale through the link function $\Phi(\cdot)$. Figure 11 panels (a)-(c) show posterior samples of the average probability that a child experiences diarrhea, fever, or cough by age. Generally, children experienced fever more often than diarrhea, and cough more often than fever. In all of these mean processes, there appears to be a local maxima before year 1. The effects
Figure 12: (a)-(c) Posterior samples of fitted weight trajectories for different subjects, $\mu(t) + (\lambda_1(t), \ldots, \lambda_5(t))^\top \eta_i + \sum_{j=1}^4 \int_T \beta_j(s, t) z_{i,j}(s) ds$ overlaid on $y_i(t_i)$ for $i = 421, 1626, 2205$.

on weight of experiencing these illnesses with all other variables held constant are shown in panels (d)-(f). Diarrhea tends to affect weight more than fever, and fever tends to affect weight more than cough. These regression functions tend to be more negative at higher ages. We speculate that the effects of diarrhea and fever affect weight more than cough because they can lead to dehydration (National Library of Medicine (US) 2022).

Finally, inference for all model components combine to estimate functions that underlie the sparsely recorded and noisy observations $y_i(t_i), i = 1, \ldots, n$. Posterior samples of $\mu(t) + (\lambda_1(t), \ldots, \lambda_5(t))^\top \eta_i + \sum_{j=1}^4 \int_T \beta_j(s, t) z_{i,j}(s) ds$ for a few children are shown in Figure 12. The model is able to fit a variety of shapes of underlying functions, with greater uncertainty at ages where the weight of children are not observed e.g. panel (b). For this dataset, the model provides reasonable fit to the observations while enabling interpretable inference with uncertainty quantification.

In Appendix F.3, we present an analysis of these data using the approach of (Crainiceanu & Goldsmith 2010) as a comparable Bayesian model. The inferential results are generally consistent with the ReMO model presented in this section with a few marked differences. In the CG model, there is no ability to propagate uncertainty in the estimated mean process and FPCs while inferring covariate effects on weight. Consequently, the empirical Bayes setup of the CG model can distort the inferences made about effect sizes, as illustrated in Section 4.3. There are discrepancies between the CG model and the ReMO model in
the magnitude of effects of season of birth and breastfeeding. In the CG model, season of birth has a relatively larger effect on weight, and breastfeeding has a relatively smaller effect on weight. To further compare the two models, we use the widely applicable information criterion (WAIC) to assess predictive performance (Vehtari et al. 2017). The ReMO model ($WAIC_{ReMO} = 69709.10$) is preferred to the CG model ($WAIC_{CG} = 69905.07$).

In Appendix F.4, we assess the fit of the ReMO model by studying estimated residuals, performing posterior predictive checks and consider reduced models by omitting covariates with seemingly small effects. We find that the model is adequate in describing individual and population level features of the observations, and the full model with all covariates considered in this section is preferred over reduced models according to WAIC.

6 Discussion

We have presented a novel approach for Bayesian FPCA through the introduction of ReMO processes. Our approach shrinks towards the orthogonality constraint needed to make meaningful inference about FPCs, while lending itself to a computationally simple Gibbs sampler. As illustrated by the data analysis of the Cebu Longitudinal Health and Nutrition Survey, our approach can flexibly model functional data, while enabling interpretable inference with coherent uncertainty quantification. We have identified several directions for future work.

In the current implementation of our approach, we condition on the number of FPCs, $K$. We could also incorporate uncertainty in $K$ and model this quantity. In order to favor a more parsimonious model, we could use a shrinkage prior, borrowing approaches from the factor analysis literature (Bhattacharya & Dunson 2011, Moran et al. 2021, Schiavon et al. 2022).

One aspect of functional data that we have not addressed in our model is the presence of phase variability, which is used to refer to the variability of the relative timing of functional features, such as extrema. Marron et al. (2015) discuss this concept and state that models
that explicitly account for this variability can be more parsimonious. In future work, we plan to extend our model for FPCA to account for phase variability, along the lines of Kneip & Ramsay (2008). One potential challenge is developing a prior model for phase that is computationally efficient and can scale to a large number of observed functions.

In this work, we have developed a prior model for near-mutually orthogonal functions that enables a computational simple approach to Bayesian FPCA. We believe that these concepts can be adapted for developing Bayesian PCA approaches for other data objects, such as surfaces and graphs. In these model extensions, the primary challenge will be the need to account for the specific intricacies of the data objects.

**Acknowledgements**

This work was supported by NIH grants RO1ES027489 and R01ES028804.
Supplemental Material: Bayesian Functional Principal Components Analysis using Relaxed Mutually Orthogonal Processes

Abstract

This is the supplemental material for the paper entitled ‘Bayesian Functional Principal Component Analysis using Relaxed Mutually Orthogonal Processes’. In Section A we present proofs for Propositions 2.1-2.6 and 3.1. In Section B we provide full implementation details for the FPCA models and GFPCA model for binary functional data. In Section C we study the sensitivity of the posterior to $\nu_\lambda$ for a simulated dataset. In Section D we reiterate the hierarchical models for FPCA and GFPCA presented in the main paper. In Section E we present an additional simulation, where data are generated from an FPCA model inferred from the Berkeley growth dataset. In Section F we provide additional details related to the data analysis of the Cebu Longitudinal Health and Nutritional Survey.

A Proofs

Proposition 2.1. Let $E := \{\lambda_1, \ldots, \lambda_K \in \mathcal{H} \mid \langle \lambda_j, \lambda_k \rangle \neq 0, \text{ for some } j \neq k\}$. For any measurable subset $E' \subseteq E$, $\lim_{\nu_\lambda \to 0} R(E') = 0$.

Proof. For any $(\lambda_1, \ldots, \lambda_k) \in E'$, $\lim_{\nu_\lambda \to 0} \exp \left\{ -\frac{1}{2\nu_\lambda} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} = 0$.

By definition of $R$,

$$\lim_{\nu_\lambda \to 0} R(E') \propto \lim_{\nu_\lambda \to 0} \int_{E'} \exp \left\{ -\frac{1}{2\nu_\lambda} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} dG(\lambda_1, \ldots, \lambda_K)$$

$$= \int_{\mathcal{H}} \lim_{\nu_\lambda \to 0} 1_{\{\lambda_1, \ldots, \lambda_K \in E'\}} \exp \left\{ -\frac{1}{2\nu_\lambda} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} dG(\lambda_1, \ldots, \lambda_K)$$

$$= \int_{\mathcal{H}} 0 dG(\lambda_1, \ldots, \lambda_K) = 0.$$  

The equality from Equation (10) to (11) follows from the Bounded Convergence Theorem, since $\exp \left\{ -\frac{1}{2\nu_\lambda} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} \leq 1$. 

31
Proposition 2.2. For any measurable set $A \subseteq \mathcal{H}$, $\lim_{\nu_{\lambda} \to \infty} R(A) = G(A)$.

Proof. For any measurable subset $A \subseteq \mathcal{H}$, $\lim_{\nu_{\lambda} \to \infty} \exp \left\{ -\frac{1}{2\nu_{\lambda}} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} = 1$ for any $\lambda_1, \ldots, \lambda_K \in A$.

By definition of $R$,

$$
\lim_{\nu_{\lambda} \to \infty} R(A) \propto \lim_{\nu_{\lambda} \to \infty} \int_A \exp \left\{ -\frac{1}{2\nu_{\lambda}} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} dG(\lambda_1, \ldots, \lambda_K) \tag{12}
$$

$$
= \int_{\mathcal{H}} \lim_{\nu_{\lambda} \to \infty} 1_{\{\lambda_1, \ldots, \lambda_K \in A\}} \exp \left\{ -\frac{1}{2\nu_{\lambda}} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} dG(\lambda_1, \ldots, \lambda_K) \tag{13}
$$

$$
= \int_{\mathcal{H}} 1_{\{\lambda_1, \ldots, \lambda_K \in A\}} dG(\lambda_1, \ldots, \lambda_K) = \int_A dG(\lambda_1, \ldots, \lambda_K) = G(A).
$$

The equality from Equation (12) to (13) follows from the Bounded Convergence Theorem, since $\exp \left\{ -\frac{1}{2\nu_{\lambda}} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} \leq 1$. Note that in Equation (12), the constant of proportionality is $\left( \lim_{\nu_{\lambda} \to \infty} \int_{\mathcal{H}} \exp \left\{ -\frac{1}{2\nu_{\lambda}} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} dG(\lambda_1, \ldots, \lambda_K) \right)^{-1} = G(\mathcal{H})^{-1} = 1$. 

Proposition 2.3. Let $E_\omega := \{\lambda_1, \ldots, \lambda_K \in \mathcal{H} \mid \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 > \omega\}$. For any $\nu_{\lambda} > 0$ and measurable subset $E'_\omega \subseteq E_\omega$ with non-zero $G(E'_\omega)$,

$$
\frac{R(E'_\omega)}{G(E'_\omega)} \leq F \exp \left( -\frac{\omega}{2\nu_{\lambda}} \right),
$$

where $F = (\mathbb{E}_G [\exp \left\{ -\frac{1}{2\nu_{\lambda}} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\}])^{-1}$ is a normalizing constant.

Proof. Note that for any $\lambda_1, \ldots, \lambda_K \in E'_\omega$, $\exp \left\{ -\frac{1}{2\nu_{\lambda}} \sum_{k=1}^{K} \sum_{j<k} \langle \lambda_j, \lambda_k \rangle^2 \right\} \leq \exp \left( -\frac{\omega}{2\nu_{\lambda}} \right)$. 

$\square$
By the definition of $R$,

$$R(E'_ω) = F \int_{E'_ω} \exp \left\{ -\frac{1}{2\nuλ} \sum_{k=1}^{K} \sum_{j<k} \langle λ_j, λ_k \rangle^2 \right\} dG(λ_1, \ldots, λ_K)$$

$$= F \int_{H} 1_{\{λ_1, \ldots, λ_K ∈ E'_ω\}} \exp \left\{ -\frac{1}{2\nuλ} \sum_{k=1}^{K} \sum_{j<k} \langle λ_j, λ_k \rangle^2 \right\} dG(λ_1, \ldots, λ_K)$$

$$≤ F \exp \left( -\frac{ω}{2νλ} \right) \int_{E'_ω} dG(λ_1, \ldots, λ_K) = F \exp \left( -\frac{ω}{2νλ} \right) G(E'_ω)$$

Since we assumed $G(E'_ω) > 0$, it follows

$$\frac{R(E'_ω)}{G(E'_ω)} \leq F \exp \left( -\frac{ω}{2νλ} \right).$$

\[\square\]

**Proposition 2.4.** Given \( \{λ_j\}_{j \neq k} \), $λ_k$ is a zero-mean Gaussian process with covariance function

$$C_{k}^{ω}(s,t) = C_k(s,t) - h_{Λ_{(-k)}}(s)(νλI_{K-1} + H_{Λ_{(-k)}})^{-1}h_{Λ_{(-k)}}(t), \text{ where} \quad (14)$$

$$h_{Λ_{(-k)}}(t) = \int_{T} C_k(s,t)Λ_{(-k)}(s)ds, \quad H_{Λ_{(-k)}} = \int_{T} \int_{T} C_k(s,s')Λ_{(-k)}(s)Λ_{(-k)}(s')^Tdsds' \quad (15)$$

with $Λ_{(-k)}(s) = (λ_1(s), \ldots, λ_{k-1}(s), λ_{k+1}(s), \ldots, λ_K(s))^T$.

**Proof.** Let $H_k$ denote the reproducing kernel Hilbert space defined by the $k^{th}$ parent GP, with Gaussian measure $G_k$. Similarly, let $H_{(-k)}$ denote the product reproducing kernel Hilbert spaces defined by all parent GPs except the $k^{th}$, with Gaussian product measure
Then, for any measurable $A \subset H_k$ and $B \subset H_{(-k)}$,

$$P_R(\lambda_k \in A, \{\lambda_j\}_{j \neq k} \in B) \propto \int_{H_k} 1_{\lambda_k \in A, \{\lambda_j\}_{j \neq k} \in B} \exp \left\{ - \frac{1}{2\nu_\lambda} \sum_{k=1}^{K} \sum_{j < k} \langle \lambda_j, \lambda_k \rangle^2 \right\} dG(\lambda_1, \ldots, \lambda_K)$$

$$= \int_{H_{(-k)}} \int_{H_k} 1_{\lambda_k \in A} \exp \left\{ - \frac{1}{2\nu_\lambda} \sum_{j \neq k} \langle \lambda_j, \lambda_k \rangle^2 \right\} dG_k(\lambda_k)$$

$$\times 1_{\{\lambda_j\}_{j \neq k} \in B} \exp \left\{ - \frac{1}{2\nu_\lambda} \sum_{j \neq k} \sum_{j' < j} \langle \lambda_j, \lambda_{j'} \rangle^2 \right\} dG_{(-k)}(\{\lambda_j\}_{j \neq k}),$$

where $R_{(-k)}$ is the ReMO measure defined using all parent GPs except the $k^{th}$. We denote $R_k(A) \propto \int_A \exp \left\{ - \frac{1}{2\nu_\lambda} \sum_{j \neq k} \langle \lambda_j, \lambda_k \rangle^2 \right\} dG_k(\lambda_k)$, which is the conditional probability measure (Resnick 2019, Section 5.8) induced by $\lambda_k|\{\lambda_j\}_{j \neq k}$, when $\lambda_1, \ldots, \lambda_K \sim R$.

To show $\lambda_k|\{\lambda_j\}_{j \neq k}$ is a Gaussian process, we will show the conditional characteristic functional $E_{R_k} \left[ \exp \left\{ i \langle U, \lambda_k \rangle \right\} |\{\lambda_j\}_{j \neq k} \right]$ coincides with the characteristic functional of a zero-mean Gaussian process with covariance function $C_*^{\nu_\lambda}(\cdot, \cdot)$, for any function $U \in H_k$ that satisfies $E_{R_k} \left[ \exp \left\{ i \langle U, \lambda_k \rangle \right\} |\{\lambda_j\}_{j \neq k} \right] < \infty$. We will make use of the lower Riemann sum, which satisfies $\frac{|T|}{m} \sum_{l=1}^{m} U(t_l^*) \lambda_k(t_l^*) \leq \langle U, \lambda_k \rangle$ for $t_l = a + l \frac{(b-a)}{m}$, $l = 0, 1, \ldots, m$, $t_l^* \in [t_{l-1}, t_l]$, $l = 1, \ldots, m$, $T = [a, b]$. In general, we use $f(t^*) = (f(t_1^*), \ldots, f(t_m^*))^\top$ to denote a vector of function evaluations on the grid $t^* = (t_1^*, \ldots, t_m^*)^\top$. 34
First, note

\[ \mathbb{E}_{G_k} \left[ \exp \left\{ - \frac{1}{2\nu^\lambda} \sum_{j \neq k} (\lambda_j, \lambda_k)^2 \right\} | \{\lambda_j\}_{j \neq k} \right] \]

\[ = \mathbb{E}_{G_k} \left[ \lim_{m \to \infty} \exp \left\{ - \frac{|T|^2}{2\nu^\lambda m^2} \lambda_j(t^*)^\top \Lambda_{-k}(t^*) \Lambda_{-k}(t^*)^\top \lambda_j(t^*) \right\} | \{\lambda_j\}_{j \neq k} \right] \]

\[ = \lim_{m \to \infty} \int_{\mathbb{R}^m} \exp \left\{ - \frac{|T|^2}{2\nu^\lambda m^2} \lambda_j(t^*)^\top \Lambda_{-k}(t^*) \Lambda_{-k}(t^*)^\top \lambda_j(t^*) \right\} \]

\[ \times (2\pi \det [C_k(t^*, t^*)])^{-\frac{1}{2}} \exp \left\{ - \frac{1}{2} \lambda_j(t^*)^\top C_k(t^*, t^*)^{-1} \lambda_j(t^*) \right\} d\lambda_k(t^*) \]

\[ = \det [C_k(t^*, t^*)]^{-\frac{1}{2}} \det \left[ C_k(t^*, t^*)^{-1} + \frac{|T|^2}{\nu^\lambda m^2} \Lambda_{-k}(t^*) \Lambda_{-k}(t^*)^\top \right]^{-\frac{1}{2}} \]  

(16)

where the interchange of expectation and the limit from Equation (16) to (17) follows from the bounded convergence theorem, since \( \exp \left\{ - \frac{|T|^2}{2\nu^\lambda m^2} \lambda_j(t^*)^\top \Lambda_{-k}(t^*) \Lambda_{-k}(t^*)^\top \lambda_j(t^*) \right\} \leq 1. \)

Similarly, note

\[ \mathbb{E}_{G_k} \left[ \exp \left\{ i(U, \lambda_k) \right\} \exp \left\{ - \frac{1}{2\nu^\lambda} \sum_{j \neq k} (\lambda_j, \lambda_k)^2 \right\} | \{\lambda_j\}_{j \neq k} \right] \]

\[ = \lim_{m \to \infty} \int_{\mathbb{R}^m} \exp \left\{ i |T| m U(t^*)^\top \lambda_k(t^*) \right\} \left( 2\pi \det [C_k(t^*, t^*)] \right)^{-\frac{1}{2}} \]

\[ \times \exp \left\{ - \frac{1}{2} \lambda_j(t^*)^\top (C_k(t^*, t^*)^{-1} + \frac{|T|^2}{\nu^\lambda m^2} \Lambda_{-k}(t^*) \Lambda_{-k}(t^*)^\top) \lambda_j(t^*) \right\} d\lambda_k(t^*) , \]

(18)

where the interchange of expectation and limit in Equation (18) follows from the bounded convergence theorem, since \( \exp \left\{ i |T| m U(t^*)^\top \lambda_k(t^*) - \frac{|T|^2}{2\nu^\lambda m^2} \lambda_j(t^*)^\top \Lambda_{-k}(t^*) \Lambda_{-k}(t^*)^\top \lambda_j(t^*) \right\} \leq 1. \)
Finally, consider

\[ \mathbb{E}_{R_k} \left[ \exp \left\{ i \langle U, \lambda_k \rangle \right\} \{\lambda_j\}_{j \neq k} \right] = \frac{\mathbb{E}_{G_k} \left[ \exp \left\{ i \langle U, \lambda_k \rangle \right\} \exp \left\{ \frac{1}{2\nu_\lambda} \sum_{j \neq k} \langle \lambda_j, \lambda_k \rangle^2 \right\} \{\lambda_j\}_{j \neq k} \right]}{\mathbb{E}_{G_k} \left[ \exp \left\{ \frac{1}{2\nu_\lambda} \sum_{j \neq k} \langle \lambda_j, \lambda_k \rangle^2 \right\} \{\lambda_j\}_{j \neq k} \right]} \]

\[ = \lim_{m \to \infty} \int_{\mathbb{R}^m} \exp \left\{ i \frac{|T|}{m} \mathbf{U}(\mathbf{t}^*)^\top \lambda_k(\mathbf{t}^*) \right\} (2\pi)^{-\frac{m}{2}} \det \left[ C_k(\mathbf{t}^*, \mathbf{t}^*)^{-1} + \frac{|T|^2}{\nu_\lambda m^2} \Lambda_{-k}(\mathbf{t}^*) \Lambda_{-k}(\mathbf{t}^*)^\top \right]^{\frac{1}{2}} \]

\[ \times \exp \left\{ -\frac{1}{2} \lambda_j(\mathbf{t}^*)^\top \left( C_k(\mathbf{t}^*, \mathbf{t}^*)^{-1} + \frac{|T|^2}{\nu_\lambda m^2} \Lambda_{-k}(\mathbf{t}^*) \Lambda_{-k}(\mathbf{t}^*)^\top \right) \lambda_j(\mathbf{t}^*) \right\} d\lambda_k(\mathbf{t}^*) \]

\[ = \lim_{m \to \infty} \exp \left\{ -\frac{1}{2} \frac{|T|^2}{m^2} \mathbf{U}(\mathbf{t}^*)^\top \left( C_k(\mathbf{t}^*, \mathbf{t}^*)^{-1} + \frac{|T|^2}{\nu_\lambda m^2} \Lambda_{-k}(\mathbf{t}^*) \Lambda_{-k}(\mathbf{t}^*)^\top \right)^{-1} \mathbf{U}(\mathbf{t}^*) \right\} \]

\[ = \lim_{m \to \infty} \exp \left\{ -\frac{1}{2} \frac{|T|^2}{m^2} \mathbf{U}(\mathbf{t}^*)^\top \left( C_k(\mathbf{t}^*, \mathbf{t}^*) - \frac{|T|}{m} C_k(\mathbf{t}^*, \mathbf{t}^*) \Lambda_{-k}(\mathbf{t}^*) \right) \right. \]

\[ \times \left( \nu_\lambda I_{K-1} + \frac{|T|^2}{m^2} \Lambda_{-k}(\mathbf{t}^*)^\top C_k(\mathbf{t}^*, \mathbf{t}^*) \Lambda_{-k}(\mathbf{t}^*) \right)^{-1} \frac{|T|}{m} \Lambda_{-k}(\mathbf{t}^*)^\top C_k(\mathbf{t}^*, \mathbf{t}^*) \mathbf{U}(\mathbf{t}^*) \right\} \]

\[ = \exp \left\{ -\frac{1}{2} \int \int \int T \mathcal{C}_k^{\nu_\lambda}(s, s') U(s) U(s') ds ds' \right\}, \]

where the matrix inversion from Equations (19) to (20) is computed via the Woodbury’s formula [Harville 1998]. In Equation (21), the characteristic functional of \( \lambda_k | \{\lambda_j\}_{j \neq k} \) coincides with the characteristic functional of a zero-mean Gaussian process with covariance function \( C_k^{\nu_\lambda}(\cdot, \cdot) \), as desired.

\( \square \)

**Proposition 2.5.** \( \lambda_k | \{\lambda_j\}_{j \neq k} \overset{d}{\to} \lambda_k^\perp | \{\lambda_j\}_{j \neq k} \) as \( \nu_\lambda \to 0 \).

**Proof.** Under the assumption that \( H \) is a positive definite matrix, it follows that \( \lim_{\nu_\lambda \to 0} (\nu_\lambda I_{K-1} + \)
\( H^{-1} = H^{-1} \). Consider,

\[
\lim_{\nu \lambda \to 0} \mathbb{E}_R k \left[ \exp \left\{ i \langle U, \lambda_k \rangle \right\} | \{ \lambda_j \}_{j \neq k} \right] = \lim_{\nu \lambda \to 0} \exp \left\{ -\frac{1}{2} \int_T \int_T C_k^{\nu \lambda} (s, s') U(s) U(s') ds ds' \right\}
\]

\[
= \exp \left\{ -\frac{1}{2} \int_T \int_T C_k (s, s') U(s) U(s') ds ds' \right\}
\]

\[
\times \lim_{\nu \lambda \to 0} \exp \left\{ \frac{1}{2} \left( \int_T U(s) h(s) ds \right)^T \left( \nu \lambda I_{K-1} + H^{-1} \right) \left( \int_T U(s') h(s') ds' \right) \right\}
\]

\[
= \exp \left\{ -\frac{1}{2} \int_T \int_T C_k (s, s') U(s) U(s') ds ds' \right\}
\]

\[
\times \exp \left\{ \frac{1}{2} \left( \int_T U(s) h(s) ds \right)^T H^{-1} \left( \int_T U(s') h(s') ds' \right) \right\}
\]

\[
= \exp \left\{ -\frac{1}{2} \int_T \int_T C_k^{+} (s, s') U(s) U(s') ds ds' \right\}.
\]

\[
\square
\]

**Proposition 2.6.** \( \langle \lambda_j, \lambda_k \rangle | \{ \lambda_j \}_{j \neq k} \overset{P}{\to} 0 \) for any \( j \neq k \) as \( \nu \lambda \to 0 \)

**Proof.** First, consider

\[
\mathbb{E}_R k \left[ \int_T \int_T | \lambda_k (s) \lambda_k (s') \lambda_j (s) \lambda_j (s') | ds ds' | \{ \lambda_j \}_{j \neq k} \right] =
\]

\[
\mathbb{E}_R k [ || \lambda_j \lambda_k ||^2_1 | \{ \lambda_j \}_{j \neq k} ]
\]

\[
= || \lambda_j ||^2_2 \mathbb{E}_R k [ || \lambda_k ||^2_2 | \{ \lambda_j \}_{j \neq k} ]
\]

\[
= || \lambda_j ||^2_2 \int_T \mathbb{E}_R k [ \lambda_k (s)^2 | \{ \lambda_j \}_{j \neq k} ] ds
\]

\[
= || \lambda_j ||^2_2 \int_T C_k^{+} (s, s) ds
\]

\[
\leq || \lambda_j ||^2_2 \int_T C_k (s, s) ds
\]
This implies that the expectation and integration operations in the left hand side of Equation (22) can be interchanged, by Fubini’s theorem.

Define $e_j$ to be a binary vector of length $K-1$ that satisfies $\lambda_j(t) = e_j^\top \lambda_{(-k)}(t)$. For any $\epsilon > 0$, by Markov’s inequality,

\[
\lim_{\nu_\lambda \to 0} P_{R_k}(\{|\langle \lambda_j, \lambda_k \rangle| > \epsilon\{|\lambda_j\}_{j \neq k}\}) \leq \lim_{\nu_\lambda \to 0} \frac{1}{\epsilon^2} \mathbb{E}_{R_k}[|\langle \lambda_j, \lambda_k \rangle|^2\{|\lambda_j\}_{j \neq k}] \\
= \lim_{\nu_\lambda \to 0} \frac{1}{\epsilon^2} \mathbb{E}_{R_k} \left[ \int_T \int_T \lambda_k(s)\lambda_k(s')\lambda_j(s)\lambda_j(s') dsds' \{|\lambda_j\}_{j \neq k}\right] \\
= \lim_{\nu_\lambda \to 0} \frac{1}{\epsilon^2} \int_T \int_T \mathbb{E}_{R_k}[\lambda_k(s)\lambda_k(s')\lambda_j(s)\lambda_j(s')\{|\lambda_j\}_{j \neq k}] dsds' \\
= \lim_{\nu_\lambda \to 0} \frac{1}{\epsilon^2} \int_T \int_T C_k(s,s')\lambda_j(s)\lambda_j(s') dsds' \\
= \lim_{\nu_\lambda \to 0} \frac{1}{\epsilon^2} \left( e_j^\top He_j - e_j^\top H(\nu_\lambda I_{K-1} + H)^{-1}He_j \right) = 0.
\]

\[\square\]

**Proposition 3.1.** $\eta_k \xrightarrow{d} (I_n - \frac{1}{n}1_n1_n^\top)N(0,I_n)$ as $\nu_\eta \to 0$.

**Proof.** First notice, $(I_n - \frac{1}{n}1_n1_n^\top)N(0,I_n) \xrightarrow{d} N(0,(I_n - \frac{1}{n}1_n1_n^\top))$, since $(I_n - \frac{1}{n}1_n1_n^\top)$ is a symmetric idempotent matrix. Next, for any $\epsilon > 0$, choose $\nu_\eta < \frac{\epsilon n^2}{\|1_n1_n^\top\|^2}$, and consider

\[
\|\text{var}(\eta_k) - (I_n - \frac{1}{n}1_n1_n^\top)\|_2 = \frac{\nu_\eta}{n\nu_\eta + n^2}\|1_n1_n^\top\|_2 < \frac{\nu_\eta}{n^2}\|1_n1_n^\top\|_2 < \epsilon.
\]

Hence, $\text{var}(\eta_k) \to (I_n - \frac{1}{n}J)$ as $\nu_\eta \to 0$. Finally, for $s \in \mathbb{R}^n$, consider the characteristic
function of $\eta_k$, 

$$
\lim_{\nu \to 0} \mathbb{E} \left[ \exp \left( i s^\top \eta_k \right) \right] = \lim_{\nu \to 0} \exp \left( - \frac{1}{2} s^\top \text{var}(\eta_k) s \right) = \exp \left( - \frac{1}{2} s^\top (I_n - \frac{1}{n} 1_n 1_n^\top) s \right),
$$

as desired.

\[\Box\]

### B MCMC Implementation

Throughout the algorithms presented in this section, we use the superscript $\text{cur}$ and $\text{can}$ to denote current and candidate states of parameters in the Markov chain, and use superscripted square brackets to denote saved parameter values. The notation $\cdot|\cdot$ is used to denote a parameter conditioned on all other parameters and the data.

For the FPCA model, the full conditional distribution of $\mu(t)$ is

$$
\mu(t)|\cdot \sim N(\mathbb{E}[\mu(t)|\cdot], \mathbb{V}[\mu(t)|\cdot]), \quad (24)
$$

$$
\mathbb{V}[\mu(t)|\cdot] = \left( C^{-1}_\mu(t,t) + \frac{1}{\sigma^2} \sum_{i=1}^n O_i^\top O_i \right)^{-1}
$$

$$
\mathbb{E}[\mu(t)|\cdot] = \frac{1}{\sigma^2} \mathbb{V}[\mu(t)|\cdot] \sum_{i=1}^n \left( O_i^\top y_i(t_i) - O_i^\top O_i \sum_{k=1}^k \eta_{i,k} \lambda_k(t) \right).
$$

The full conditional distribution of $\lambda_k(t), k = 1, \ldots, K$ is

$$
\lambda_k(t)|\cdot \sim N(\mathbb{E}[\lambda_k(t)|\cdot], \mathbb{V}[\lambda_k(t)|\cdot]), \quad (25)
$$

$$
\mathbb{V}[\lambda_k(t)|\cdot] = \left( C^{-1}_k(t,t) + \frac{1}{\nu_k} W \Lambda^\top(-k)(t) \Lambda(-k)(t) W + \frac{1}{\sigma^2} \sum_{i=1}^n \eta_{i,k}^2 O_i^\top O_i \right)^{-1}
$$

$$
\mathbb{E}[\lambda_k(t)|\cdot] = \frac{1}{\sigma^2} \mathbb{V}[\lambda_k(t)|\cdot] \sum_{i=1}^n \left( \eta_{i,k} O_i^\top y_i(t_i) - O_i^\top O_i \mu(t) - O_i^\top O_i \sum_{j \neq k} \eta_{i,j} \lambda_j(t) \right).
$$
The full conditional distribution of \( \eta_k, k = 1, \ldots, K \) is
\[
\eta_k | - \sim N(\mathbb{E}[\eta_k | -], \mathbb{V}[\eta_k | -]), \tag{26}
\]
\[
\mathbb{V}[\eta_k | -] = \left( \text{diag}(\lambda_k(t)^\top O_i^\top O_i \lambda_k(t)) \right)_{i=1}^n + \frac{1}{\psi_k} \left( I_n + \frac{1}{\nu} 1_n 1_n^\top \right)^{-1},
\]
\[
\mathbb{E}[\eta_k | -] = \frac{1}{\sigma^2} \mathbb{V}[\eta_k | -] \left[ (y_i(t_i) - O_i \mu(t) - O_i \sum_{j \neq k} \eta_{i,j} \lambda_j(t))^\top O_i \lambda_k(t) \right]_{i=1}^n,
\]
where \( \text{diag}(\cdot)_{i=1}^n \) denotes a diagonal matrix with diagonal elements given by the argument, and \([\cdot]_{i=1}^n\) is a vector with elements given by the argument.

The full conditional distribution of \( \psi_k, k = 1, \ldots, K \) is
\[
\psi_k | - \sim \text{inverse-gamma}(\alpha_\eta + \frac{n}{2}, \beta_\eta + \frac{1}{2} \eta_k^\top (I_n + \frac{1}{\nu} 1_n 1_n^\top)^{-1} \eta_k), \tag{27}
\]

The full conditional distribution of \( \sigma^2 \) is
\[
\sigma^2 | - \sim \text{inverse-gamma} \left( \alpha_\sigma + \frac{1}{2} \sum_{i=1}^n m_i, \beta_\sigma + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^{m_i} (y_i(t_{i,j}) - \mu(t_{i,j}) + \sum_{k=1}^K \eta_{i,k} \lambda_k(t_{i,j}))^2 \right). \tag{28}
\]

A Metropolis-within-Gibbs algorithm for our GFPCA approach for binary functional data is presented in Algorithm 2. The latent variable formulation of the probit model leads to a simple step for sampling the latent processes \( z_{i}^*, i = 1, \ldots, n \) as described in Albert & Chib (1993). Conditioning on the latent processes, the sampling of the model parameters is similar to the FPCA setting.

C Sensitivity Analysis For \( \nu_\lambda \)

While the sensitivity of the prior on \( \nu_\lambda \) is discussed at length in the main paper, we focus on the sensitivity of the posterior on \( \nu_\lambda \) for a simulated dataset in this section. Figure 13 panel
Algorithm 1: Metropolis-within-Gibbs for FPCA

1: Randomly initialize current states of all parameters using their prior distribution
2: for iter = 1 : N do
3: Draw $\mu^{\text{cur}}(t) \sim \pi(\mu(t))$ from Equation (24)
4: Propose $l^{\text{can}} \sim N(l^{\text{cur}}, \Sigma^{\text{prop}})$
5: Compute $\alpha_{\mu} := \frac{\pi(\mu^{\text{cur}}(t)|l^{\text{can}})\pi(\sigma^{\text{can}})}{\pi(\mu^{\text{cur}}(t)|l^{\text{cur}})\pi(\sigma^{\text{cur}})}$
6: If unif(0, 1) < $\alpha_{\mu}$, set $l^{\text{cur}} = l^{\text{can}}$
7: Propose $\tau^{2,\text{can}} \sim N(\tau^{2,\text{cur}})$
8: Compute $\alpha_{\tau^{2}} := \frac{\pi(\tau^{2,\text{cur}}|l^{\text{can}})\pi(\tau^{2,\text{can}})}{\pi(\tau^{2,\text{cur}}|l^{\text{cur}})\pi(\tau^{2,\text{can}})}$
9: If unif(0, 1) < $\alpha_{\tau^{2}}$, set $\tau^{2,\text{cur}} = \tau^{2,\text{can}}$
10: for k in 1 : K do
11: Draw $\lambda^{\text{cur}}(t) \sim \pi(\lambda_k(t))$ from Equation (25)
12: Propose $l^{\text{can}} \sim N(l^{\text{cur}}, \Sigma^{\text{prop}})$
13: Compute $\alpha_{\lambda_k} := \frac{\pi(\lambda^{\text{cur}}(t)|l^{\text{can}})\pi(\sigma^{\text{can}})}{\pi(\lambda^{\text{cur}}(t)|l^{\text{cur}})\pi(\sigma^{\text{cur}})}$
14: If unif(0, 1) < $\alpha_{\lambda_k}$, set $l^{\text{cur}} = l^{\text{can}}$
15: Propose $\tau^{2,\text{can}} \sim N(\tau^{2,\text{cur}})$
16: Compute $\alpha_{\tau^{2}} := \frac{\pi(\tau^{2,\text{cur}}|l^{\text{can}})\pi(\tau^{2,\text{can}})}{\pi(\tau^{2,\text{cur}}|l^{\text{cur}})\pi(\tau^{2,\text{can}})}$
17: If unif(0, 1) < $\alpha_{\tau^{2}}$, set $\tau^{2,\text{cur}} = \tau^{2,\text{can}}$
18: Draw $\eta^{\text{cur}} \sim \pi(\eta_k)$ from Equation (26)
19: Draw $\psi^{\text{cur}} \sim \pi(\psi_k)$ from Equation (27)
20: Draw $\sigma^{2,\text{cur}} \sim \pi(\sigma^2)$ from Equation (28)
21: Save

$\mu^{[\text{iter}]}(t), l^{[\text{iter}], \tau^{2,\text{iter}}, \lambda^{[\text{iter}], \eta^{[\text{iter}], \psi^{[\text{iter}]}, k}]_{k=1}, \sigma^{2,\text{iter}}}$

(a) shows a simulated dataset that was generated according to the FPCA observation model presented in the main paper, with $K = 2$. Panel (b) shows the 2 FPCs that underlie the observations in panel (a). These are specified as the orthogonal functions $\lambda_1(t) := \sin(t)$ and $\lambda_2(t) := \frac{1}{2}\sin(2t)$. To study the effects of $\nu_k$ on posterior inference, we run several MCMC implementations of our model at different values of $\nu_k = 100, 1, 0.01, 1e-04, 1e-06, 1e-08$. During these different implementations we fix $K = 5$, and the initialization and the seed for the psuedo-random number generator to simulate MCMC samples from the target posterior. Panels (d)-(i) of Figure 13 show trace plots of 10,000 posterior samples of $||\lambda_k||_2$, $k = 1, \ldots, 5$, with the true values of $||\lambda_k||_2$, $k = 1, 2$ shown as black horizontal lines.
Algorithm 2 Metropolis-within-Gibbs for GFPCA

1: Randomly initialize current states of all parameters using their prior distribution
2: for iter = 1 : N do
3:     for i = 1 : n do
4:         for j = 1 : m_i do
5:             Draw \( z_i^{* \text{cur}}(t_{i,j}) \) | ~

\[
\begin{align*}
\text{truncated-normal}_{[0,\infty)}(\mu_{\text{cur}}(t_{i,j}) + \sum_{k=1}^{K} \eta_{i,k}^{\text{cur}} \lambda_{k}^{\text{cur}}(t_{i,j}), 1), \quad & z_i(t_{i,j}) = 1, \\
\text{truncated-normal}_{(-\infty,0)}(\mu_{\text{cur}}(t_{i,j}) + \sum_{k=1}^{K} \eta_{i,k}^{\text{cur}} \lambda_{k}^{\text{cur}}(t_{i,j}), 1), \quad & z_i(t_{i,j}) = 0.
\end{align*}
\]

6: \( \mu_{\text{cur}}(t), \mu_{\text{cur}}, \tau_{\mu_{\text{cur}}}^{2 \text{cur}}; \lambda_{k}(t)_{\text{cur}}, \lambda_{k}^{2 \text{cur}}, \eta_{k}^{\text{cur}}, \psi_{k}^{\text{cur}} \}_{k=1}^{K} \) can be sampled using steps 3 - 19 of Algorithm 1, replacing \( y_i := z_i^{*}, \ i = 1, \ldots, n \) and \( \sigma^2 := 1 \)
7: Save \( \mu_{\text{iter}}(t), \mu_{\text{iter}}, \tau_{\mu_{\text{iter}}}^{2 \text{iter}}; \lambda_{k}(t)_{\text{iter}}, \lambda_{k}^{2 \text{iter}}, \eta_{k}^{\text{iter}}, \psi_{k}^{\text{iter}} \}_{k=1}^{K} \) := \( \mu_{\text{cur}}(t), \mu_{\text{cur}}, \tau_{\mu_{\text{cur}}}^{2 \text{cur}}; \lambda_{k}(t)_{\text{cur}}, \lambda_{k}^{2 \text{cur}}, \eta_{k}^{\text{cur}}, \psi_{k}^{\text{cur}} \}_{k=1}^{K}

In the setting where \( \nu_\lambda \) is large, as in panels (d)\&(e), the model is not able to identify the FPCs, and the inferred FPCs split the true factors as indicated by the norms of the MCMC samples failing to converge. At the other extreme, panels (h)\&(i) show posterior samples of FPCs that mix slowly, and take a large number of iterations to converge to a posterior mode. Panel (c) shows the effective sample size (ESS) of the posterior samples of \( \|\lambda_k\|_2, \ k = 1, 2 \) after the first 5,000 MCMC iterations. The ESS of \( \|\lambda_k\|_2, \ k = 1, 2 \) is small for \( \nu_\lambda = 100 \) and 1 because of lack of identifiability, while the ESS for \( \nu_\lambda = 1e - 08 \) is small because of slow mixing. A value of \( \nu_\lambda = 1e - 04 \) balances these two extremes and has a relatively large ESS.
Figure 13: (a) Simulated observations. (b) Latent FPCs that underlie the observations. (c) Effective sample size of $\|\lambda_k\|_2$, $k = 1, 2$ for different values of $\nu_{\lambda}$. Trace plots of $\|\lambda_k\|_2$, $k = 1, \ldots, 5$ for (d) $\nu_{\lambda} = 100$, (e) $\nu_{\lambda} = 1$, (f) $\nu_{\lambda} = 0.01$, (g) $\nu_{\lambda} = 1e^{-04}$, (h) $\nu_{\lambda} = 1e^{-06}$, (i) $\nu_{\lambda} = 1e^{-08}$. 
### D Hierarchical Models

For FPCA our hierarchical model is specified as follows,

\[
y_i | \mu(t), \lambda_1(t), \ldots, \lambda_K(t), \eta_i \overset{\text{iid}}{\sim} N_m(O_i \mu(t) + O_i(\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i, \sigma^2 I_m), \quad i = 1, \ldots, n
\]

\[
\mu | l_\mu, \tau^2_\mu \sim GP(0, C_\mu(\cdot, \cdot)), \quad l^2_\mu \sim \text{inverse-gamma}(\alpha_\mu, \beta_\mu), \quad \tau^2_\mu \sim \text{half-normal}(\gamma_\mu)
\]

\[
\lambda_k | l_k, \tau^2_k, \Lambda_{(-k)} \sim GP(0, C_k^{\mu_\lambda}(\cdot, \cdot)), \quad \eta_k | \psi_k \sim N(0, \psi_k(I_n + \frac{1}{\nu_\eta} 1_n 1_n^\top)^{-1}), \quad k = 1, \ldots, K
\]

\[
l^2_k \sim \text{inverse-gamma}(\alpha_\lambda, \beta_\lambda), \quad \tau^2_k \sim \text{half-normal}(\gamma_\lambda), \quad \psi_k \sim \text{inverse-gamma}(\alpha_\eta, \beta_\eta), \quad k = 1, \ldots, K
\]

\[
\sigma^2 \sim \text{inverse-gamma}(\alpha_\sigma, \beta_\sigma).
\]

The hierarchical GFPCA extension for binary functional data is

\[
z_i(t_i) = \begin{cases} 1, & z^*_i(t_i) > 0, \\ 0, & \text{otherwise}, \quad i = 1, \ldots, n, \end{cases}
\]

\[
z^*_i(t_i) | \mu(t), \lambda_1(t), \ldots, \lambda_K(t), \eta_i \overset{\text{iid}}{\sim} N_m(O_i \mu(t) + O_i(\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i, I_m), \quad i = 1, \ldots, n
\]

\[
\mu | l_\mu, \tau^2_\mu \sim GP(0, C_\mu(\cdot, \cdot)), \quad l^2_\mu \sim \text{inverse-gamma}(\alpha_\mu, \beta_\mu), \quad \tau^2_\mu \sim \text{half-normal}(\gamma_\mu)
\]

\[
\lambda_k | l_k, \tau^2_k, \Lambda_{(-k)} \sim GP(0, C_k^{\mu_\lambda}(\cdot, \cdot)), \quad \eta_k | \psi_k \sim N(0, \psi_k(I_n + \frac{1}{\nu_\eta} 1_n 1_n^\top)^{-1}), \quad k = 1, \ldots, K
\]

\[
l^2_k \sim \text{inverse-gamma}(\alpha_\lambda, \beta_\lambda), \quad \tau^2_k \sim \text{half-normal}(\gamma_\lambda), \quad \psi_k \sim \text{inverse-gamma}(\alpha_\eta, \beta_\eta), \quad k = 1, \ldots, K.
\]

### E Additional Simulated Example

In this simulated example we generate observations that replicate the variability in data from the Berkeley growth study, described in [Ramsay & Silverman 2005](https://www.springer.com/gp/book/9780387954568). Panel (a) of Figure 14 displays the height in centimeters of the 39 male children in the study. Using the `fdapace` R package we inferred the components of an FPCA model that implements
Figure 14: (a) Observed growth curves for $n = 39$ male subjects in the Berkeley growth study. (b) MISE of estimated functions for different FPCA methods.

The methodology of Yao et al. (2005), under the following observation model

$$y_i(t) = \mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i + \epsilon_i(t), \; i = 1, \ldots, n.$$  \hspace{1cm} (29)

The ages $t$ correspond to quarterly measurements in the first year of age, annual measurements from age two to age eight, then biannual measurements from age 8 to 18. The \texttt{fdapace} package selected $K = 4$, so that the estimated fraction of variance explained by the FPCs is 95\%. In this simulation, we sample independent Gaussian FPC scores and additive error with variance estimated from \texttt{fdapace}, and plug estimates of the mean and FPCs into the above observation model to generate functional datasets with variability similar to that of Figure 14(a). Using the simulated observations, we measure estimation performance of various methods using mean integrated squared error $MISE = \frac{1}{n} \sum_{i=1}^{n} \int_T (\hat{f}_i(s) - f_i(s))^2 ds$, where $\hat{f}_i$ is an estimate of $f_i = \mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i$.

For the ReMO model, we use the posterior mean of $\mu(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i$ to produce an estimate of $f_i$ based on 1,000 MCMC samples. In terms of other FPCA methods, we use PACE, Newton, and CG as in the main paper. In all of these methods, we specify $K = 4$, and we replicate the entire data generating mechanism 100 times for comparison. Panel (b) of Figure 14 shows MISE results. Newton is excluded from the
figure because the MISE was orders of magnitudes larger than the other three methods. The MISE computed using ReMO, PACE, and CG estimates are all comparable, with ReMO slightly edging out the competitors. Even in this realistic simulation setting under model misspecification, our FPCA model using ReMO processes performs well compared to competing methods.

F Additional Information for the Analysis of the Cebu Longitudinal Health and Nutrition Survey Dataset

F.1 Hierarchical Model Formulation

In Section 5 of the main paper, we motivated a flexible model for weight of young children while accounting for scalar and longitudinal covariates. The observation model is restated as follows,

\[ y_i(t_i) = O_i \mu(t) + O_i (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i + O_i \sum_{j=1}^{4} \int_{T} \beta_j(s, t) z_{i,j}(s) ds + \epsilon_i(t_i) \quad (30) \]

\[ \eta_i = \Theta x_i + \xi_i \quad (31) \]

Similar to the hierarchical model for FPCA presented in section [D], we specify the following model components:

\[ \mu|l_\mu, \tau^2_\mu \sim GP(0, C_\mu(\cdot, \cdot)), \ l^2_\mu \sim \text{inverse-gamma}(\alpha_\mu, \beta_\mu), \ \tau^2_\mu \sim \text{half-normal}(\gamma_\mu) \]

\[ \lambda_k|l_k, \tau^2_k, \Lambda(-k) \sim GP(0, C^k(\cdot, \cdot)), \ k = 1, \ldots, K \]

\[ l^2_k \sim \text{inverse-gamma}(\alpha_\lambda, \beta_\lambda), \ \tau^2_k \sim \text{half-normal}(\gamma_\lambda) \ k = 1, \ldots, K \]

\[ \epsilon_i(t_i)|\sigma^2 \sim iid N(0, I_m), \ i = 1 \ldots, n, \ \sigma^2 \sim \text{inverse-gamma}(\alpha_\sigma, \beta_\sigma). \]
We incorporate scalar covariates, \(x_i\), through the FPC scores, \(\eta_i = Bx_i + \xi_i\). For these model parameters, we assign

\[
\Theta_{k,q}|\psi_k \sim N(0, \psi_k), \; q = 1, \ldots, 4, \; \xi_k|\psi_k \sim N(0, \psi_k(I_n + \frac{1}{\nu_\psi}1_n1_n^\top)^{-1}), \; k = 1, \ldots, K
\]

\[
\psi_k \sim \text{inverse-gamma}(\alpha_\xi, \beta_\xi), k = 1, \ldots, K
\]

As noted in the main paper, the longitudinal covariates are sparsely recorded. Based on initial EDA of the data, the illness indicator covariates, \(z_{i,2}, z_{i,3}, z_{i,4}, i = 1, \ldots, n\), do not appear to have much structured variability across subjects. For model based imputation of these covariates, we model the probability of illness at time \(s\) as

\[
p(z_{i,j}(s) = 1) = \Phi(\mu_{z^j}(s)), \; i = 1, \ldots, n, j = 2, 3, 4.
\]

Alternatively, we do suspect that there is structured variability in the breastfeeding indicator \(z_{i,1}\), given typical breastfeeding patterns and biological constraints. We model \(P(z_{i,1}(t) = 1) = \Phi(\mu_{z^1}(s) + (\lambda_1^z(s), \ldots, \lambda_K^z(s))\eta_i^z)\) as in the GFPCA setting of our model. For these components, we assign

\[
\mu_{z^j}(s)|l_{\mu z^j}, \tau_{\mu z^j}^2 \sim GP(0, C_{\mu z^j}(\cdot, \cdot)), j = 1, \ldots, 4
\]

\[
l_{\mu z^j}^2 \sim \text{inverse-gamma}(\alpha_{\mu z^j}, \beta_{\mu z^j}), \; \tau_{\mu z^j}^2 \sim \text{half-normal}(\gamma_{\mu z^j}), j = 1, \ldots, 4
\]

\[
\lambda_k^z|l_k^z, \tau_k^z, \Lambda_{-k}^z \sim GP(0, C_{\lambda_k^z}(\cdot, \cdot)), \; \eta_k^z|\psi_k \sim N(0, \psi_k(I_n + \frac{1}{\nu_\psi}1_n1_n^\top)^{-1}),
\]

\[
l_k^z \sim \text{inverse-gamma}(\alpha_{\lambda z^j}, \beta_{\lambda z^j}), \; \tau_k^z \sim \text{half-normal}(\gamma_{\lambda z^j}), \; \psi_k \sim \text{inverse-gamma}(\alpha_{\eta z^j}, \beta_{\eta z^j}),
\]

\[k = 1, \ldots, K.
\]

For grid points where the covariates are not observed, \(s_{i,j}^{\text{miss}}\), \(j = 1, \ldots, 4\), values are imputed for \(z_{i,j}(s_{i,j}^{\text{miss}})\) based on the above models. These values are sampled throughout the MCMC algorithm to account for the uncertainty due to the missing data.

Finally, we can specify the model components for the functional linear model component \([\text{Ramsay \& Dalzell 1991}]. Each \beta(s,t)\) is represented via a basis expansion. For the historical linear model component for the breastfeeding indicator, we use the tent basis described in
We denote these basis elements with $V_{p,1}(s,t)$, $p = 1, \ldots, P$, so that $\int_T \beta_1(s,t) z_{i,1}(s) ds = \sum_{p=1}^{P_1} \rho_{p,1} \int_T V_{p,1}(s,t) z_{i,1}(s) ds$, where the integrals are approximated numerically based on the imputed breastfeeding indicator covariate. For the illness covariates, the integral form equates to $\int_T \beta_j(s,t) z_{i,j}(s) ds = \beta_j(t) z_{i,j}(t)$, $j = 2, 3, 4$ in the concurrent linear model [Hastie & Tibshirani 1993]. We represent each with a b-spline basis, $\beta_j(t) = \sum_{p=1}^{P_j} \rho_{p,j} V_{p,j}(t)$. Inference for these model components is carried out through inference on the basis coefficients for which we specify

$$
\rho_{p,j} \sim N(0, \tau_{p,j}^2), \ p = 1, \ldots, P_j, \ j = 1, \ldots, 4.
$$

**F.2 Additional Figures for Posterior Inference**

In the main paper, the inferential results for the analysis of the Cebu Longitudinal Health and Nutrition Survey are reported with $K = 5$ and $K_z = 3$. These numbers of FPCs were based on exploratory data analysis. Figure 15 displays scree plots of norms of inferred FPCs for modeling weight and breastfeeding trajectories based on a larger number of components. In both of these cases, the first few FPCs are significantly larger in magnitude compared to trailing FPCs. We chose to present $K = 5$ in the main paper, which corresponds to describing 99.99% of structured variability in weight (estimated through PACE), and $K_z = 3$, since the magnitude of latter FPCs is negligible. As noted in the discussion, future work for this modelling framework includes allowing the number of components to be flexible so they are not fixed *a priori*, and to incorporate shrinkage for latter components to encourage parsimony.

As discussed in the main paper, model based imputation is needed for the longitudinal covariates in order to compute the integral equations of the functional linear model components of the model. Figure 16 shows the patterns of missingness by age for the breastfeeding and illness indicator functions.

Figure 17 shows FPC scores for weight and breastfeeding status. In panel (a), the FPC
scores of weight appear much more normal, relative to the FPC scores of breastfeeding status shown in panel (b). We investigate atypical observations for breastfeeding in Figure 18. Outliers in terms of the second FPC score are shown in panels (a)-(c). In general, it looks like these outliers correspond to the situation where a child was reported not to have breastfed erratically at an early age. Panels (d)-(f) show outliers in terms of the third FPC score. These outliers correspond to the situation where a child was reported not to have breastfed sometime after age 1, but continued to breastfeed after age 1.5. In addition to these outliers, the non-normal structure of the FPC scores is replicated in exploratory data analyses presented in Section F.3.
In this section, we analyze the Cebu dataset with methods discussed in Crainiceanu & Goldsmith (2010), termed ‘CG’ in the main paper, which we compare to the results from the ReMO model. The observation model is given by Equations (30) & (31). In line with their empirical Bayes setup, $\mu(t)$ is estimated using the cross sectional mean of the data, and $\lambda_1, \ldots, \lambda_K$ are estimated from the mean-centered data using a penalized spline based approach. Additionally, we complete longitudinal covariates using model based imputation, with the generative model

$$P(z_{i,1}(t) = 1) = \logit^{-1}(\mu_{z_{i1}} + (\lambda_{z_{i1}}(t), \ldots \lambda_{z_Kz_{iK}}(t))^T \eta_i^z),$$

$$P(z_{i,j}(t) = 1) = \logit^{-1}(\mu_{z_{ij}}), \ j = 2, 3, 4,$$

where model components are estimated using the `slfpc` R package.

The mean of the response and FPCs are fixed at point estimates, $\hat{\mu}(t), \hat{\lambda}_1(t), \ldots, \hat{\lambda}_K(t)$ in subsequent modeling. Remaining model parameters are assigned conjugate prior distributions and posterior inference is carried out through Gibbs sampling. As suggested by Crainiceanu & Goldsmith (2010), a priori $\Theta_{k,q} \overset{\text{ind}}{\sim} N(0, \hat{\phi}_k)$ and $\xi \overset{\text{ind}}{\sim} N(0, \hat{\phi}_k)$, where the prior variances are eigenvalues of the estimated covariance matrix that determine the estimated FPCs. For the regression coefficients between the scalar covariates and the FPC scores, we assume independent, identically distributed standard normal priors. We use the same bases discussed in the main paper to simplify the integral equations in the functional linear model components that govern the relationships between the longitudinal covariates and weight. Basis coefficients are given normal conjugate priors, consistent with the ReMO model.

Figure 19 panels (a) and (b) show the estimated mean process and FPCs. These estimated functions are similar in shape and magnitude to the FPCs presented in the
main paper. Panels (d)-(f) of this figure show the inferred effects of the scalar covariates by displaying posterior samples of $(\hat{\lambda}_1(t), \ldots, \hat{\lambda}_K(t))\hat{\Theta}_q$, $q = 1, \ldots, 4$. For $q = 1, 2, 3$ the resulting functions match the results presented in the main paper. The function for $q = 4$ in the main paper is inferred to be closer to zero with a considerable amount of uncertainty. The discrepancy between these two results could be from conditioning on estimated FPCs in the empirical Bayes setting.

Figure 20 panels (a)-(c) visualize estimated FPCs used to describe the structure variability in breastfeeding status estimated from the \texttt{slfpcas} R package. These FPCs seem to capture if a child is breastfed, and when they stopped breastfeeding. Panels (d) of this figure shows estimated FPC scores, $\hat{\eta}_i^z$, $i = 1, \ldots, n$. As in Section F.2, the FPC scores have a non-Gaussian structure. Panel (e)-(h) display $\| \int_{s - s'}^t \hat{\beta}_1(s') ds' \|_2$, for different values of $s'$. The general shape of inferred functions is consistent with those presented in the main paper, however, the functions in panels (e)-(h) are slightly smaller in magnitude. As shown in panel (f), there is a diminishing effect of prolonged breastfeeding.

Figure 21 panel (a) displays the estimated average probability of a child experiencing the illnesses by age. As in the model presented in the main paper, all of these estimated functions have a prominent peak at early ages. The estimated regression functions, $\beta_j(t)$, $j = 2, 3, 4$, shown in panels (b)-(d), match their corresponding interpretations in the main paper. It appears that experiencing diarrhea and fever have a more negative effect on weight than experiencing cough.

The inferences made in this section are generally consistent with the inferences presented in the main paper. We use the widely applicable information criterion (WAIC) to compare the predictive ability of the model presented in this section and the ReMO model (Vehtari et al. 2017). WAIC is $\text{WAIC}_{\text{ReMO}} = 69709.10$ for the ReMO model and $\text{WAIC}_{\text{CG}} = 69905.07$ for the CG model. The difference in WAICs is $\text{WAIC}_{\text{ReMO}} - \text{WAIC}_{\text{CG}} = -195.97$ with standard error 114.04. Using this criteria, the ReMO model is preferred in terms of predictive accuracy compared to the CG model.
F.4 Model Fit Assessment

In this section, we assess the fit of the ReMO model presented in the main paper.

Let \( f_i(t) = \mu_i(t) + (\lambda_1(t), \ldots, \lambda_K(t))^\top \eta_i + \sum_{j=1}^4 \int_T \beta_j(s, t) z_{ij}(s) ds \) denote the function that underlies the sparse and noisy observation \( y_i(t_i) \), \( f_i^{[\text{iter}]}(t) \) denote posterior MCMC samples, and let \( \hat{f}_i(t) \) denote the posterior mean computed from MCMC samples. Residuals are estimated as \( \hat{\epsilon}_i(t_i) = y_i(t_i) - \hat{f}_i(t_i) \). Figure 23 shows a boxplot of estimated residuals for all subjects, \( i = 1, \ldots, n \), by age. Visually, it appears that there is no mean trend across ages, and there does not appear to be substantial heteroskedasticity.

We further assess the fit of the model using posterior predictive checks (Gelman et al. 2013). Under the assumed model, posterior predictive samples are generated via \( y_i^{\text{pred},[\text{iter}]}(t_i) \sim N(f_i^{[\text{iter}]}(t), \sigma_i^{[\text{iter}]}I_m) \). Quantities of interest, \( T \), based on the posterior predictive samples are compared to the same quantities computed using the actual observations. We consider the following quantities of interest used to study the fit of our model,

\[
T_{\text{mag},y_i} = \|y_i\|_2, \quad i = 1, \ldots, n \\
T_{\text{mean},(y_1,\ldots,y_n)}(t) = \frac{1}{n} \sum_{i=1}^n y_i(t), \quad t = 0, \frac{2}{24}, \ldots, 2 \\
T_{\text{var},(y_1,\ldots,y_n)}(k) = \text{eig}_k \left( \text{cov}(y(s), y(t)) \right), \quad k = 1, \ldots, K,
\]

where \( \text{eig}_k \) denotes the \( k^{\text{th}} \) largest eigenvalue of the sample covariance matrix based on \( y_1, \ldots, y_n \). The quantity \( T_{\text{mag}} \) can be used to assess how well the model captures the magnitude of individual subjects, \( T_{\text{mean}} \) can be used to assess how well the model captures the mean trend across subjects, and \( T_{\text{var}} \) can be used to assess how well the modes of variability are captured across subjects.

Figure 24 panels (a)-(f) show histograms of \( T_{\text{mag},y_i} \) based on posterior predictive draws for different subjects, where the red line represents \( T_{\text{mag},y_i} \) based on actual observed values. Panels (a)-(c) represent random samples of subjects in the population, while panels (d)-(f) represent subjects having posterior predictive mean of \( T_{\text{mag},y_i} \) furthest from the observed
Table 1: WAIC for full versus reduced models. In all cases the full model is preferred.

| covariate omitted | WAIC      | difference (full - reduced) | standard error (difference in WAIC) |
|-------------------|-----------|----------------------------|-------------------------------------|
| cough             | 69727.66  | -18.40                     | 26.41                               |
| season            | 69719.45  | -10.19                     | 33.76                               |
| strata            | 69718.37  | -9.11                      | 26.41                               |

value. Even in the worst case scenario, the predictions based on the model cover the observed value. Panel (g) of this figure shows a histogram of the absolute difference between the mean of $T_{mag,y_i}$ based on posterior predictive samples and the observed value. The absolute difference is mostly concentrated near zero. Figure 25 panels (a)-(f) show histograms of $T_{mean,(y_1,...,y_n)}$ based on posterior predictive draws at different times $t = 0, \ldots, 2$, where the red line represents $T_{mean,(y_1,...,y_n)}$ based on actual observed values. These values are computed cross-sectionally by removing missing values. At each time, it appears that the observed value of $T_{mean,(y_1,...,y_n)}$ is covered by the values computed using posterior predictive samples. Figure 26 panels (a)-(f) show histograms of $T_{var,(y_1,...,y_n)}$ based on posterior predictive draws for different eigenvalues $k = 1, \ldots, 5$, where the red line represents $T_{var,(y_1,...,y_n)}$ based on actual observed values. The sample covariance is computed by using pairwise complete observations. For each eigenvalue the observed value of $T_{var,(y_1,...,y_n)}$ is covered by the values computed using posterior predictive samples. Based on these posterior predictive checks, it is apparent that the model is adequate in describing these individual and population level features of the observations.

In the main paper, cough, season, and strata are noted as the covariates with the smallest effect on the response. We consider reducing the model presented in the main paper by omitting these covariates. In order to assess fit of reduced models, compared to a full model with all covariates, we use WAIC. Results are presented in Table 1. For each case, the full model is preferred.
F.5 MCMC Trace Plots

The inferential results of the main paper are based on MCMC samples. In this section we plot thinned MCMC samples to diagnose convergence. The diagnostic plots for all parameters do not indicate that there is an issue with MCMC convergence. Figure 27 displays trace plots related to the inferential results of Figures 7&12 of the main paper. Figure 28 displays trace plots related to the inferential results of Figure 8 of the main paper. Figure 29 displays trace plots related to the inferential results of Figures 9&10 of the main paper. Figure 30 displays trace plots related to the inferential results of Figures 11 of the main paper.

References

Adair, L., Popkin, B. M., VanDerslice, J., Akin, J., Guilkey, D., Black, R., Briscoe, J. & Flieger, W. (1993), ‘Growth dynamics during the first two years of life: A prospective study in the Philippines’, European Journal of Clinical Nutrition 47(1), 42–51.

Adair, L. S. & Guilkey, D. K. (1997), ‘Age-specific determinants of stunting in Filipino children’, The Journal of Nutrition 127(2), 314–320.

Adair, L. S., Popkin, B. M., Akin, J. S., Guilkey, D. K., Gultiano, S., Borja, J., Perez, L., Kuzawa, C. W., McDade, T. & Hindin, M. J. (2011), ‘Cohort profile: The Cebu longitudinal health and nutrition survey’, International Journal of Epidemiology 40(3), 619–625.

Albert, J. H. & Chib, S. (1993), ‘Bayesian analysis of binary and polychotomous response data’, Journal of the American Statistical Association 88(422), 669–679.

Andrieu, C. & Thoms, J. (2008), ‘A tutorial on adaptive MCMC’, Statistics and Computing 18(4), 343–373.

Behseta, S., Kass, R. E. & Wallstrom, G. L. (2005), ‘Hierarchical models for assessing variability among functions’, Biometrika 92(2), 419–434.
Bhattacharya, A. & Dunson, D. B. (2011), ‘Sparse Bayesian infinite factor models’, *Biometrika* **98**(2), 291–306.

Brumback, B. A. & Rice, J. A. (1998), ‘Smoothing spline models for the analysis of nested and crossed samples of curves’, *Journal of the American Statistical Association* **93**(443), 961–976.

Cardot, H. (2000), ‘Nonparametric estimation of smoothed principal components analysis of sampled noisy functions’, *Journal of Nonparametric Statistics* **12**(4), 503–538.

Casey, C. E., Neifert, M. R., Seacat, J. M. & Neville, M. C. (1986), ‘Nutrient intake by breast-fed infants during the first five days after birth’, *American Journal of Diseases of Children* **140**(9), 933–936.

Castro, P. E., Lawton, W. H. & Sylvestre, E. (1986), ‘Principal modes of variation for processes with continuous sample curves’, *Technometrics* **28**(4), 329–337.

Crainiceanu, C. M. & Goldsmith, A. J. (2010), ‘Bayesian functional data analysis using WinBUGS’, *Journal of Statistical Software* **32**(11), 1–33.

Dewey, K. G. (1998), ‘Growth characteristics of breast-fed compared to formula-fed infants’, *Neonatology* **74**(2), 94–105.

Duan, L. L., Young, A. L., Nishimura, A. & Dunson, D. B. (2019), ‘Bayesian constraint relaxation’, *Biometrika* **107**(1), 191–204.

Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis*, Chapman and Hall/CRC.

Ghosh, J. & Dunson, D. B. (2009), ‘Default prior distributions and efficient posterior computation in Bayesian factor analysis’, *Journal of Computational and Graphical Statistics* **18**(2), 306–320.
Goldsmith, J., Zipunnikov, V. & Schrack, J. (2015), ‘Generalized multilevel function-on-scalar regression and principal component analysis’, *Biometrics* 71(2), 344–353.

Hall, P., Müller, H.-G. & Yao, F. (2008), ‘Modelling sparse generalized longitudinal observations with latent Gaussian processes’, *Journal of the Royal Statistical Society: Series B* 70(4), 703–723.

Harville, D. A. (1998), *Matrix Algebra from a Statistician’s Perspective*, Springer.

Hastie, T. & Tibshirani, R. (1993), ‘Varying-coefficient models’, *Journal of the Royal Statistical Society: Series B* 55(4), 757–779.

James, G. M., Hastie, T. J. & Sugar, C. A. (2000), ‘Principal component models for sparse functional data’, *Biometrika* 87(3), 587–602.

Jauch, M., Hoff, P. D. & Dunson, D. B. (2021), ‘Monte Carlo simulation on the Stiefel manifold via polar expansion’, *Journal of Computational and Graphical Statistics* 30(3), 622–631.

Jiang, L., Zhong, Y., Elrod, C., Natarajan, L., Knight, R. & Thompson, W. K. (2020), ‘Bayestime: Bayesian functional principal components for sparse longitudinal data’, arXiv:2012.00579.

Kneip, A. & Ramsay, J. O. (2008), ‘Combining registration and fitting for functional models’, *Journal of the American Statistical Association* 103(483), 1155–1165.

Malfait, N. & Ramsay, J. O. (2003), ‘The historical functional linear model’, *Canadian Journal of Statistics* 31(2), 115–128.

Marcus, M. B. & Shepp, L. A. (1972), Sample behavior of Gaussian processes, in ‘Proceedings of the Sixth Berkeley Symposium on Mathematics, Statistics and Probability’, Vol. 2, pp. 423–421.

Marron, J. S., Ramsay, J. O., Sangalli, L. M. & Srivastava, A. (2015), ‘Functional data analysis of amplitude and phase variation’, *Statistical Science* 30(4), 468–484.
Montagna, S., Tokdar, S. T., Neelon, B. & Dunson, D. B. (2012), ‘Bayesian latent factor regression for functional and longitudinal data’, *Biometrics* **68**(4), 1064–1073.

Moran, K. R., Dunson, D., Wheeler, M. W. & Herring, A. H. (2021), ‘Bayesian joint modeling of chemical structure and dose response curves’, *The Annals of Applied Statistics* **15**(3), 1405–1430.

National Library of Medicine (US) (2022), ‘Dehydration’, MedlinePlus. URL: [https://medlineplus.gov/dehydration.html](https://medlineplus.gov/dehydration.html).

Nolan, T. H., Goldsmith, J. & Ruppert, D. (2021), ‘Bayesian functional principal components analysis via variational message passing’, arXiv:2104.00645.

Peng, J. & 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.

Plumlee, M. & Joseph, V. R. (2018), ‘Orthogonal Gaussian process models’, *Statistica Sinica* **28**(2), 601–619.

Poworoznek, E., Ferrari, F. & Dunson, D. (2021), ‘Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching’, arXiv:2107.13783.

Ramsay, J. O. & Dalzell, C. (1991), ‘Some tools for functional data analysis’, *Journal of the Royal Statistical Society: Series B* **53**(3), 539–561.

Ramsay, J. & Silverman, B. (2005), *Functional Data Analysis*, Springer Series in Statistics, Springer.

Rasmussen, C. E. & Williams, C. K. (2006), *Gaussian Processes for Machine Learning*, Vol. 2, MIT press.

Resnick, S. (2019), *A Probability Path*, Springer.
Rice, J. A. & Silverman, B. W. (1991), ‘Estimating the mean and covariance structure nonparametrically when the data are curves’, *Journal of the Royal Statistical Society: Series B* 53(1), 233–243.

Rice, J. A. & Wu, C. O. (2001), ‘Nonparametric mixed effects models for unequally sampled noisy curves’, *Biometrics* 57(1), 253–259.

Schiavon, L., Canale, A. & Dunson, D. B. (2022), ‘Generalized infinite factorization models’, *Biometrika*. asab056.

Suarez, A. J. & Ghosal, S. (2017), ‘Bayesian estimation of principal components for functional data’, *Bayesian Analysis* 12(2), 311 – 333.

van der Linde, A. (2008), ‘Variational Bayesian functional PCA’, *Computational Statistics & Data Analysis* 53, 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.

van Zanten, J. & van der Vaart, A. (2008), Reproducing kernel Hilbert spaces of Gaussian priors, in ‘Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh’, Institute of Mathematical Statistics.

Vehtari, A., Gelman, A. & Gabry, J. (2017), ‘Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC’, *Statistics and Computing* 27(5), 1413–1432.

Wrobel, J., Zipunnikov, V., Schrack, J. & Goldsmith, J. (2019), ‘Registration for exponential family functional data’, *Biometrics* 75(1), 48–57.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005), ‘Functional data analysis for sparse longitudinal data’, *Journal of the American Statistical Association* 100(470), 577–590.

Zhong, R., Liu, S., Li, H. & Zhang, J. (2021), ‘Sparse logistic functional principal component analysis for binary data’, arXiv:2109.08009.
Figure 17: Scatter plot matrices of posterior means of FPC scores for (a) $\xi$ (weight) and (b) $\eta$ (breastfeeding).
Figure 18: (a)-(f) Posterior samples of fitted atypical breastfeeding trajectories, $\Phi(\mu_{z_i}(t) + (\lambda_{z_1}(t), \lambda_{z_2}(t), \lambda_{z_3}(t))^T \eta_z^i)$ overlaid on $z_{i,1}(t_i)$ for $i = 212, 2782, 2809, 1516, 1844, 2701$.

Figure 19: Inferred parameters related to FPCA of weight and scalar covariates using the CG model: (a) $\hat{\mu}(t)$. (b) $\sqrt{\hat{\psi}_1 \hat{\lambda}_1}, \ldots, \sqrt{\hat{\psi}_K \hat{\lambda}_K}$. (c) Posterior means of $\xi_{i,k}$, $i = 1, \ldots, n$; $k = 1, \ldots, K$. (d)-(g) Posterior samples of $(\hat{\lambda}_1(t), \ldots, \hat{\lambda}_K(t))^{\Theta_q}$, $q = 1, \ldots, 4$. 

60
Figure 20: Inference for parameters related to modeling breastfeeding status on weight using the CG model: (a)-(c) logit $-1(\hat{\mu}_z - \text{sd}(\hat{\eta}_{i,k}, \ldots, \hat{\eta}_{n,k}) \lambda_k(t)$ for $k = 1, 2, 3$. (d) FPC scores, $\eta_{i,k}, i = 1, \ldots, n, k = 1, 2, 3$. (e)-(h) Posterior samples of $\int_T \beta_1 s_s ds' s' = 0, 0.5, 1, 2$. (f) Boxplots of posterior samples of $\| \int_T \beta_1 s_s ds' \|_2$ for different values of $s'$.

Figure 21: Inferred functions related to modelling the effects of illness on weight using the CG model: (a) logit $-1(\hat{\mu}_z(t))$. (b)-(d) Posterior samples of $\beta_j(t)$ for $j = 2, 3, 4$. 

61
Figure 22: Inferred weight trajectories for different subjects using the CG model: (a)-(c) Posterior samples of $\hat{\mu}(t_i) + (\hat{\lambda}_1(t_i), \ldots, \hat{\lambda}_K(t_i))^\top \hat{\eta}_i + O_i \sum_{j=1}^4 \int_T \beta_j(s, t) z_{i,j}(s) ds$, overlaid on $y_i(t_i), i = 421, 1626, 2205$.

Figure 23: Boxplots of residuals by age.
Figure 24: $T_{\text{mag},y_i}$ based on posterior predictive samples (histogram) and observation (red lines) for random subjects (a)-(c) and subjects whose predicted values were furthest from the observed values (d)-(f). (g) the absolute difference between the mean of $T_{\text{mag},y_i}$ based on posterior predictive samples and the observed value for all subjects.
Figure 25: $T_{\text{mean},(y_1,...,y_n)}$ based on posterior predictive samples (histogram) and observation (red lines) for (a) $t = 0$ years, (b) $t = 2/24$ years, ..., (m) $t = 2$ years.
Figure 26: $T_{\text{var}, (y_1, ..., y_n)}$ based on posterior predictive samples (histogram) and observation (red lines) for (a) $k = 1$, (b) $k = 2$, ..., (e) $k = 5$. 
Figure 27: Trace plots related to the FPCA results presented Figures 7 & 12 in the main paper.
(\lambda_1(t_1), \ldots, \lambda_5(t_1))^\top \Theta_q \text{ for } q = 1, \ldots, 4

(\lambda_1(t_6), \ldots, \lambda_5(t_6))^\top \Theta_q \text{ for } q = 1, \ldots, 4

(\lambda_1(t_{13}), \ldots, \lambda_5(t_{13}))^\top \Theta_q \text{ for } q = 1, \ldots, 4

Figure 28: Trace plots related to FPC score regression onto scalar covariates presented in Figure 8 in the main paper.
\[ \Phi(\mu_z(t) + (\lambda_z^1(t), \lambda_z^2(t), \lambda_z^3(t))^\top \eta_{2087}) \] for \( t = 0, 1, 2 \)

Figure 29: Trace plots related to breastfeeding status interpolation and historical regression presented in Figures 9&10 in the main paper.
Figure 30: Trace plots related to illness status interpolation and concurrent regression presented in Figure 11 in the main paper.