Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–27

Bayesian Functional Principal Components Analysis via Variational Message Passing with Multilevel Extensions*

Tui H. Nolan[†,§,¶] Jeff Goldsmith[‡], and David Ruppert[§,‖]

Abstract. Standard approaches for functional principal components analysis rely on an eigendecomposition of a smoothed covariance surface in order to extract the orthonormal eigenfunctions representing the major modes of variation in a set of functional data. This approach can be a computationally intensive procedure, especially in the presence of large datasets with irregular observations. In this article, we develop a variational Bayesian approach, which aims to determine the Karhunen-Loève decomposition directly without smoothing and estimating a covariance surface. More specifically, we incorporate the notion of variational message passing over a factor graph because it removes the need for rederiving approximate posterior density functions if there is a change in the model. Instead, model changes are handled by changing specific computational units, known as fragments, within the factor graph – we demonstrate this with an extension to multilevel functional data. Indeed, this is the first article to address a functional data model via variational message passing. Our approach introduces three new fragments that are necessary for Bayesian functional principal components analysis. We present the computational details, a set of simulations for assessing the accuracy and speed of the variational message passing algorithm and an application to United States temperature data.

Keywords: nonparametric regression, Kullback-Liebler divergence, functional principal component scores, mean field.

MSC2020 subject classifications: Primary 60K35, 60K35.

arXiv: 2104.00645

*Tui H. Nolan's research was supported by a Fulbright scholarship, an American Australian Association scholarship and a Roberta Sykes scholarship. Jeff Goldsmith's research was supported by Award R01NS097432 from the National Institute of Neurological Disorders and Stroke (NINDS) and by Award R01AG062401 from the National Institute of Aging. David Ruppert's research was supported by the National Science Foundation grant AST-1814840.

[†]MRC Biostatistics Unit, University of Cambridge, East Forvie Building, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SR, tn352@cam.ac.uk

[‡]Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th St. NY, NY 10032, ajg2202@cumc.columbia.edu

[§]Department of Statistics and Data Science, Cornell University, 1198 Comstock Hall, 129 Garden Ave., Ithaca, NY 14853, dr24@cornell.edu

[¶]Also affiliated with School of Mathematical and Physical Sciences, University of Technology Sydney and formerly affiliated with School of Operations Research and Information Engineering, Cornell University for the majority of this work.

[‖]Also affiliated with School of Operations Research and Information Engineering, Cornell University.

 https://doi.org/10.1214/23-BA1393

1 Introduction

Functional principal components analysis (FPCA) is the methodological extension of classical principal components analysis (PCA) to functional data. The advantages of using FPCA for functional data are derived from analogous advantages that PCA affords for multivariate data analysis. For instance, PCA in the multivariate data setting is used to reduce dimensionality and identify the major modes of variation of the data set. The modes of variation are determined by the eigenvectors of the sample covariance matrix of the data set, while dimension reduction is achieved by identifying the eigenvectors that maximize variation in the data. In the functional setting, response curves are interpreted as independent realizations of an underlying stochastic process. A covariance operator and its eigenfunctions play the analogous role that the covariance matrix and its eigenvectors play in the multivariate data setting. By identifying the eigenfunctions with the largest eigenvalues, one can reduce the dimensionality of the entire data set by approximating each curve as a linear combination of a finite set of eigenfunctions.

There are technical issues that arise in the functional setting that are not present for multivariate data. Without loss of generality, we will assume that the domain of the functional curves is $[0, 1]$. In addition, the curves are only observed at discrete, irregular points over this interval. Therefore, approaches that are used in PCA require modifications to extend to the functional framework. In FPCA, we often rely on nonparametric regression to smooth the eigenfunctions and employ an appropriate step to ensure that they are orthonormal on $L^2([0, 1])$.

There have been numerous developments in FPCA methodology throughout the statistical literature. A thorough introduction to the statistical framework and applications can be found in Ramsay and Silverman (2005, Chapter 8) and Wang et al. (2016, Section 2). Much of this work mirrors the eigendecomposition approach to PCA, in that an eigenbasis is obtained from a covariance surface. Of particular interest for our analysis, Yao et al. (2005) focused on the case of sparsely observed functional data, and estimate principal component scores through conditional expectations, while Di et al. (2009) extended FPCA to multilevel functional data, extracting within and between subject sources of variability.

Meanwhile, other approaches have built on or are similar to the probabilistic PCA framework that was introduced by Tipping and Bishop (1999) and Bishop (1999). Rather than first obtaining eigenfunctions from a smoothed covariance surface and then estimating scores, all quantities are considered unknown and are estimated jointly. James et al. (2000) used an expectation maximization algorithm for estimation and inference in the context of sparsely observed curves. Variational Bayes for FPCA was introduced by van der Linde (2008) via a generative model with a factorized approximation of the full posterior density function.

In standard versions of FPCA, the covariance function is determined through bivariate smoothing of the raw covariances. Eigenfunctions and eigenvalues are then determined from the smoothed covariance function. Finally, the scores are estimated from the covariance function, eigenfunctions and eigenvalues via best linear unbiased prediction (Yao et al., 2005). Although such an approach is built upon a coherent sequence

of conditional steps, complex bivariate smoothing for estimation of the covariance function requires storage of large covariance matrices for dense functional data. When there are few if any overlapping pairs of observations in sparse, irregular functional data, it is hard to estimate a covariance and smooth it. The key advantage in various probabilistic approaches is that the covariance function is not estimated (van der Linde, 2008; Goldsmith et al., 2015; Goldsmith and Schwartz, 2017), meaning that complex bivariate smoothing is not required. Here, one could determine eigenfunctions and scores by either maximizing a likelihood, as in James et al. (2000), or by taking a Bayesian approach by specifiying suitable priors. In the latter case, the eigenfunctions and eigenvalues are computed directly as part of a Bayesian hierarchical model. Furthermore, it is unnecessary to compute or store large covariance matrices so that direct estimation of eigenfunctions is straightforward. For these reasons, we pursue a Bayesian approach to FPCA.

Although there have been numerous contributions to Bayesian implementations of FPCA, we argue that there are additional considerations that should be addressed. First, MCMC modeling of FPCA can be a computationally expensive procedure. For instance, Goldsmith et al. (2015) uses Hamiltonian Monte Carlo sampling via Stan (Stan Development Team, 2020) to perform FPCA as part of a generalized multilevel function-on-scalar regression model. Binary functional data indicating physical activity or inactivity for 600 subjects over 5 days were analyzed using 5000 iterations of the sampler, with the total computation time being 10 days. Second, current versions of variational Bayes for FPCA, despite being a much faster computational alternative, are difficult to extend to more complex likelihood specifications. In particular, multilevel extensions are of key interest for our application to US temperature data.

Minka (2005) presents a unifying view of approximate Bayesian inference under a message passing framework that relies on the notion of messages passed between nodes of a factor graph. Mean field variational Bayes (MFVB) (Ormerod and Wand, 2010; Blei et al., 2017) can be incorporated into this framework through an alternate scheme known as variational message passing (VMP) (Winn and Bishop, 2005). Wand (2017) introduced computational units, known as fragments, that compartmentalize the algebraic derivations that are necessary for approximate Bayesian inference in VMP. The notion of fragments within a factor graph is essential for efficient extensions of variational Bayes-based FPCA to arbitrarily large statistical models. In this article, we demonstrate this directly by extending a VMP-based Bayesian FPCA model to its multilevel counterpart, while only deriving one extra fragment.

It is important to note that the MFVB and VMP algorithms are based on the same optimisation problem. Therefore, they converge to identical posterior distributions. Previous analysis (Nolan, 2020) has shown that MFVB algorithms tend to converge faster than VMP algorithms, but only on the order of seconds. However, this does not take into account the time saved in mathematical derivations and coding through the VMP approach (Wand, 2017). In particular, VMP is easier to incorporate in a coordinated modeling framework.

In this article, we propose an FPCA extension of the VMP framework for variational Bayesian inference set out in Wand (2017). Our novel methodology includes the

introduction of three fragments that are necessary for computing approximate posterior density functions via variational inference, as well as a sequence of post-processing steps for estimating the orthonormal eigenfunctions. Section 2 gives an overview of FPCA and introduces the Bayesian hierarchical model. We provide an introduction to variational Bayesian inference in Section 3, with an overview of VMP in Section 3.1. The utility of the VMP approach is made evident in Section 4, where we extend the variational Bayesian algorithm to the multilevel setting. In Section 5, we outline the post-VMP steps that are required for producing orthonormal eigenfunctions. Simulations, including speed and accuracy comparisons with MCMC algorithms, are presented in Section 6, and an application to United States temperature data is provided in Section 7.

1.1 Matrix Algebraic Background

We define the vec and vech operators, which are well-established (e.g. Gentle, 2007). For a $d_1 \times d_2$ matrix, the vec operator concatenates the columns of the matrix from left to right. For a $d_1 \times d_1$ matrix, the vech operator concatenates the columns of the matrix after removing the above diagonal elements. For example, suppose that $\boldsymbol{A} = \begin{bmatrix} (2,-3)^\intercal & (-1,1)^\intercal \end{bmatrix}$. Then $\text{vec}(\boldsymbol{A}) = (2,-3,-1,1)^\intercal$ and $\text{vech}(\boldsymbol{A}) = (2,-3,1)^\intercal$. For a $d^2 \times 1$ vector $\boldsymbol{a}$, $\text{vec}^{-1}(\boldsymbol{a})$ is the $d \times d$ matrix such that $\text{vec}\{\text{vec}^{-1}(\boldsymbol{a})\} = \boldsymbol{a}$. Additionally, the matrix $\boldsymbol{D}_d$ is the duplication matrix of order d, and it is such that $\boldsymbol{D}_d \,\text{vech}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A})$ for a $d \times d$ symmetric matrix $\boldsymbol{A}$. Furthermore, $\boldsymbol{D}_d^+ \equiv (\boldsymbol{D}_d^\intercal \boldsymbol{D}_d)^{-1} \boldsymbol{D}_d^\intercal$ is the Moore-Penrose inverse of $\boldsymbol{D}_d$, where $\boldsymbol{D}_d^+ \,\text{vec}(\boldsymbol{A}) = \text{vech}(\boldsymbol{A})$.

For a set of d matrices $\{\boldsymbol{M}_i\}_{i=1,\ldots,d}$, we define:

$$\underset{i=1,\ldots,d}{\text{stack}}(\boldsymbol{M}_i) \equiv \begin{bmatrix} \boldsymbol{M}_1 \\ \vdots \\ \boldsymbol{M}_d \end{bmatrix} \quad \text{and} \quad \underset{i=1,\ldots,d}{\text{blockdiag}}(\boldsymbol{M}_i) \equiv \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{M}_2 & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{M}_d \end{bmatrix},$$

with the first of these definitions requiring that each $\boldsymbol{M}_i$ has the same number of columns.

2 Functional Principal Components Analysis

Consider a set of random realizations $x_1, \ldots, x_n \in L^2[0,1]$ of a smooth Gaussian process $x(t)$. We will assume the existence of a continuous mean function $\mu = \mathbb{E}\, x_i$ and continuous covariance surface $\sigma(t,s) = \mathbb{E}[\{x_i(t) - \mu(t)\}\{x_i(s) - \mu(s)\}]$, $i = 1, \ldots, n$. Then, the covariance operator Σ of x_i is defined as $(\Sigma g)(t) \equiv \int_0^1 \sigma(t,s) g(s) ds$, $g \in L^2[0,1]$. From Mercer's Theorem, the spectral decomposition of Σ satisfies $\sigma(s,t) = \sum_{l=1}^{\infty} \gamma_l \; \psi_l(s) \; \psi_l(t)$, where the γ_l are the eigenvalues of Σ in descending order and ψ_l are the corresponding orthonormal eigenfunctions. The Karhunen-Loève decomposition is the basis for the FPCA expansion (Yao et al., 2005):

$$x_i(t) = \mu(t) + \sum_{l=1}^{\infty} \zeta_{il} \psi_l(t), \quad i = 1, \ldots, n, \tag{2.1}$$

where $\zeta_{il} = \int_0^1 \{x_i(t) - \mu(t)\}\psi_l(t)dt$ are the principal component scores. The ζ_{il} are independent across i and uncorrelated across l, with $\mathbb{E}(\zeta_{il}) = 0$ and $\mathbb{V}\text{ar}(\zeta_{il}) = \gamma_l$.

Expansion (2.1) facilitates dimension reduction by providing a best approximation for each curve $x_1, \ldots, x_n$ in terms of the truncated sums involving the first L orthonormal eigenfunctions $\psi_1, \ldots, \psi_L$. That is, for any choice of L orthonormal eigenfunctions $f_1, \ldots, f_L$, the minimum of $\sum_{i=1}^n ||x_i - \mu - \sum_{l=1}^L \langle x_i - \mu, f_l\rangle f_l||^2$ is achieved for $f_l = \psi_l$, $l = 1, \ldots, L$, where $||\cdot||$ denotes the L^2 norm and $\langle\cdot,\cdot\rangle$ denotes the L^2 inner product. For this reason, we define the best estimate of x_i as

$$\widehat{x}_i(t) \equiv \mu(t) + \sum_{l=1}^{L} \zeta_{il}\ \psi_l(t), \quad i = 1, \ldots, n. \tag{2.2}$$

For the remainder of this article, we assume that all eigenvalues of the covariance operator have multiplicity one. In addition, issues of identifiability are always present when one attempts to infer eigenfunctions or eigenvectors. However, choosing one eigenfunction over its opposite sign has no effect on the resulting fits, although one choice of sign may provide more natural interpretation of the eigenfunction. Here, we simply assume that the signs of the orthonormal eigenfunctions $\psi_1, \ldots, \psi_L$ are such that if $\widehat{\psi}_l$ is an estimator of ψ_l, then $\langle \psi_l, \widehat{\psi}_l\rangle > 0$.

Expansions similar to (2.2) are also possible, where

$$\widehat{x}_i(t) \equiv \mu(t) + \sum_{l=1}^{L} z_{il}\ h_l(t), \quad i = 1, \ldots, n, \tag{2.3}$$

where z_{il} are Gaussian random variables that are correlated across l, but remain independent across i, and the h_l are not orthonormal. Theorem 2.1 shows that an orthogonal decomposition of the resulting basis functions and scores is sufficient for establishing the appropriate estimates (2.2) from (2.3). Its proof is provided in Appendix A.1 (Nolan et al., 2023).

Theorem 2.1. *Given the decomposition in (2.3), there exists a unique set of orthonormal eigenfunctions $\psi_1, \ldots, \psi_L$ and an uncorrelated set of scores $\zeta_{i1}, \ldots, \zeta_{iL}$, $i = 1, \ldots, n$, such that $\widehat{x}_i(t) = \mu(t) + \sum_{l=1}^L \zeta_{il}\ \psi_l(t)$.*

Theorem 2.1 motivates estimation of the Karhunen-Loève decomposition directly to infer the eigenfunctions and scores. In this approach, all components of the Karhunen-Loève decomposition are viewed as unknown so that scores and eigenfunctions are estimated jointly. The other class of methods use covariance decompositions to obtain the eigenfunctions and subsequently estimate the scores given the eigenfunctions using the Karhunen-Loève decomposition (e.g. Yao et al., 2005; Di et al., 2009; Xiao et al., 2016). There are several advantages in the former method in that it does not require estimation or smoothing of a large covariance and can more directly handle sparse or irregular functional data.

2.1 Bayesian Model Construction

In practice, the curves $x_1, \ldots, x_n$ are indirectly observed as noisy observations at irregular, discrete points in time. Let the set of design points for the ith curve be summarized by the vector $\boldsymbol{t}_i \equiv (t_{i1}, \ldots, t_{in_i})^\intercal$ and the observations for the ith curve, $x_i(t)$, by the vector $\boldsymbol{x}_i \equiv \{x_i(t_{i1}) + \epsilon_{i1}, \ldots, x_i(t_{in_i}) + \epsilon_{in_i}\}^\intercal$, where n_i is the number of observations on the ith curve and ϵ_{ij} are i.i.d. noise terms with $\mathbb{E}(\epsilon_{ij}) = 0$ and $\mathbb{V}\text{ar}(\epsilon_{ij}) = \sigma_\epsilon^2$. The finite decomposition in (2.2) takes the form:

$$\boldsymbol{x}_i = \boldsymbol{\mu}_i + \sum_{l=1}^{L} \zeta_{il}\boldsymbol{\psi}_{il} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{2.4}$$

where $\boldsymbol{\mu}_i \equiv \{\mu(t_{i1}), \ldots, \mu(t_{in_i})\}^\intercal$, $\boldsymbol{\psi}_{il} \equiv \{\psi_l(t_{i1}), \ldots, \psi_l(t_{in_i})\}^\intercal$, for $l = 1, \ldots, L$, and $\boldsymbol{\epsilon}_i \equiv (\epsilon_{i1}, \ldots, \epsilon_{in_i})^\intercal$ is the vector of measurement errors for the observations on curve $x_i(t)$.

We model continuous curves from discrete observations via nonparametric regression (Ruppert et al., 2003, 2009), using the mixed model-based penalized spline basis function representation, as in Durbán et al. (2005). The representation for the mean function and the FPCA eigenfunctions are: $\mu(t) \approx \beta_{\mu,0} + \beta_{\mu,1}t + \sum_{k=1}^{K} u_{\mu,k}z_k(t)$ and $\psi_l(t) \approx \beta_{\psi_l,0} + \beta_{\psi_l,1}t + \sum_{k=1}^{K} u_{\psi_l,k}z_k(t)$, for $l = 1, \ldots, L$ where $\{z_k(\cdot)\}_{1\leq k\leq K}$ is a suitable set of basis functions. Splines and wavelet families are the most common choices for the z_k. In our simulations, we use O'Sullivan penalized splines, which are similar to P-splines, but have the advantage of being a reparameterization of smoothing splines that is convenient for a Bayesian or mixed model representation (Wand and Ormerod, 2008). In addition, the mixed model representation admits a diagonal penalty matrix and opting for cubic O'Sullivan penalized splines permits natural boundary conditions, where the second and third derivatives of the approximated nonlinear curve are zero.

In order to avoid notational clutter, we incorporate the following definitions: $\boldsymbol{\beta}_\mu \equiv (\beta_{\mu,0}, \beta_{\mu,1})^\intercal$, $\boldsymbol{u}_\mu \equiv (u_{\mu,1}, \ldots, u_{\mu,K})^\intercal$, $\boldsymbol{\nu}_\mu \equiv (\boldsymbol{\beta}_\mu^\intercal, \boldsymbol{u}_\mu^\intercal)^\intercal$; and $\boldsymbol{\beta}_{\psi_l} \equiv (\beta_{\psi_l,0}, \beta_{\psi_1,1})^\intercal$, $\boldsymbol{u}_{\psi_l} \equiv (u_{\psi_l,1}, \ldots, u_{\psi_l,K})^\intercal$, $\boldsymbol{\nu}_{\psi_l} \equiv (\boldsymbol{\beta}_{\psi_l}^\intercal, \boldsymbol{u}_{\psi_l}^\intercal)^\intercal$ for $l = 1, \ldots, L$. Then simple derivations that stem from (2.4) show that the vector of observations on each of the response curves satisfies the representation $\boldsymbol{x}_i = \boldsymbol{C}_i(\boldsymbol{\nu}_\mu + \sum_{l=1}^{L} \zeta_{il}\boldsymbol{\nu}_{\psi_l}) + \boldsymbol{\epsilon}_i$, where

$$\boldsymbol{C}_i \equiv \begin{bmatrix} 1 & t_{i1} & z_1(t_{i1}) & \ldots & z_K(t_{i1}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{in_i} & z_1(t_{in_i}) & \ldots & z_K(t_{in_i}) \end{bmatrix}. \tag{2.5}$$

In addition, we define $\boldsymbol{x} \equiv (\boldsymbol{x}_1^\intercal, \ldots, \boldsymbol{x}_n^\intercal)^\intercal$, $\boldsymbol{\nu} \equiv (\boldsymbol{\nu}_\mu^\intercal, \boldsymbol{\nu}_{\psi_1}^\intercal, \ldots, \boldsymbol{\nu}_{\psi_L}^\intercal)^\intercal$ and $\boldsymbol{\zeta}_i \equiv (\zeta_{i1}, \ldots, \zeta_{iL})^\intercal$.

Next, we present the Bayesian FPCA Gaussian response model:

$$\boldsymbol{x}_i|\boldsymbol{\nu}, \boldsymbol{\zeta}_i, \sigma_\epsilon^2 \overset{\text{ind.}}{\sim} \text{N}\left\{\boldsymbol{C}_i\left(\boldsymbol{\nu}_\mu + \sum_{l=1}^{L}\zeta_{il}\boldsymbol{\nu}_{\psi_l}\right), \sigma_\epsilon^2\boldsymbol{I}_{n_i}\right\}, \quad \boldsymbol{\zeta}_i \overset{\text{ind.}}{\sim} \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\zeta_i}), \quad i = 1, \ldots, n,$$

$$\begin{bmatrix}\boldsymbol{\nu}_\mu\\ \boldsymbol{\nu}_{\psi_l}\end{bmatrix} \Big| \sigma^2_\mu, \sigma^2_{\psi_l} \overset{\text{ind.}}{\sim} \text{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_\mu & \mathbf{O}^\intercal\\ \mathbf{O} & \boldsymbol{\Sigma}_{\psi_l}\end{bmatrix}\right), \quad \sigma^2_{\psi_l}|a_{\psi_l} \overset{\text{ind.}}{\sim} \text{Inverse}-\chi^2(1, 1/a_{\psi_l}),$$
$$a_{\psi_l} \overset{\text{ind.}}{\sim} \text{Inverse}-\chi^2(1, 1/A), \quad l = 1, \ldots, L,$$
$$\sigma^2_\mu|a_\mu \sim \text{Inverse}-\chi^2(1, 1/a_\mu), \quad a_\mu \sim \text{Inverse}-\chi^2(1, 1/A),$$
$$\sigma^2_\epsilon|a_\epsilon \sim \text{Inverse}-\chi^2(1, 1/a_\epsilon), \quad a_\epsilon \sim \text{Inverse}-\chi^2(1, 1/A), \tag{2.6}$$

where

$$\boldsymbol{\Sigma}_\mu \equiv \begin{bmatrix}\sigma^2_\beta \boldsymbol{I}_2 & \mathbf{O}^\intercal\\ \mathbf{O} & \sigma^2_\mu \boldsymbol{I}_K\end{bmatrix}, \quad \boldsymbol{\Sigma}_{\psi_l} \equiv \begin{bmatrix}\sigma^2_\beta \boldsymbol{I}_2 & \mathbf{O}^\intercal\\ \mathbf{O} & \sigma^2_{\psi_l} \boldsymbol{I}_K\end{bmatrix}, \quad l = 1, \ldots, L, \tag{2.7}$$

and $\sigma^2_\beta > 0$, $A > 0$ are the model hyperparameters. Note that the iterated inverse-χ^2 distributional specification on σ^2_ϵ, which involves an inverse-χ^2 prior specification on the auxiliary variable a_ϵ, is equivalent to $\sigma^2_\epsilon \sim \text{Half-Cauchy}(A)$. This auxiliary variable-based hierarchical construction facilitates arbitrarily non-informative priors on standard deviation parameters (Gelman, 2006). Similar comments also apply to the iterated inverse-χ^2 distributional specifications for σ^2_μ and $\sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}$. Other prior specifications, such as half-t priors on standard deviation parameters or inverse gamma priors on variance parameters, were analysed in Maestrini and Wand (2021). These distributional specifications can also be introduced into model (2.6), by replacing the iterated inverse chi-squared fragments in Figure 1 with the appropriate fragments from Maestrini and Wand (2021). The multivariate standard Gaussian prior on each vector of scores is a common specification in probabilistic PCA and its functional extensions (Tipping and Bishop, 1999; van der Linde, 2008; Goldsmith et al., 2015).

3 Variational Bayesian Inference

In keeping with the theme of this article, we will explain variational Bayesian inference and its extensions to VMP in the context of the Bayesian FPCA model (2.6). For an in-depth introduction to variational Bayesian inference, see Ormerod and Wand (2010) and Blei et al. (2017). See Minka (2005) and Wand (2017) for expositions on VMP.

Full Bayesian inference for the parameter set $\boldsymbol{\nu}, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n, \sigma^2_\epsilon, a_\epsilon, \sigma^2_\mu, a_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}$ and $a_{\psi_1}, \ldots, a_{\psi_L}$ requires the determination of the posterior density function $p(\boldsymbol{\nu}, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n, \sigma^2_\epsilon, a_\epsilon, \sigma^2_\mu, a_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}, a_{\psi_1}, \ldots, a_{\psi_L}|\boldsymbol{x})$, but it is analytically intractable. The standard approach for overcoming this deficiency is to employ MCMC approaches. However, MCMC simulations are very slow for model (2.6), even for moderate dimensions of $\boldsymbol{\nu}$, which depends on the number of eigenfunctions (L) and O'Sullivan penalized spline basis functions (K).

Alternatively, variational approximate inference for model (2.6) involves the mean field restriction:

$$p(\boldsymbol{\nu}, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n, \sigma^2_\epsilon, a_\epsilon, \sigma^2_\mu, a_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}, a_{\psi_1}, \ldots, a_{\psi_L}|\boldsymbol{x}) \approx$$
$$\left\{\prod_{i=1}^n q(\boldsymbol{\zeta}_i)\right\} q(\boldsymbol{\nu}) q(\sigma^2_\epsilon) q(a_\epsilon) q(\sigma^2_\mu) q(a_\mu) \left\{\prod_{l=1}^L q(\sigma^2_{\psi_l}) q(a_{\psi_l})\right\}, \tag{3.1}$$

where each q represents an approximate density function. The q-density functions are selected to minimize the Kullback-Leibler divergence of the left-hand side of (3.1) from its right-hand side. The approximation in (3.1) is based on assuming posterior independence between global parameters (spline coefficients for the mean curve and the eigenfunctions) and response curve-specific parameters (the scores), incorporating the notion of *asymptotic independence* between regression coefficients and variance parameters (Menictas and Wand, 2013, Section 3.1), and induced factorizations based on graph theoretic results (Bishop, 2006, Section 10.2.5). The parameter vectors that define each of the q-density functions are interrelated and are updated by a coordinate ascent algorithm (Ormerod and Wand, 2010, Algorithm 1). However, the resulting parameter vector updates are problem-specific and must be rederived if there is a change to the model.

3.1 Variational Message Passing

VMP is an alternate computational framework for variational Bayesian inference with a mean field product restriction. The VMP infrastructure is a factor graph representation of the Bayesian model. Wand (2017) advocates for the use of fragments, a sub-graph of a factor graph, as a means of compartmentalizing the algebra and computer coding required for variational Bayesian inference. Posterior density estimation is achieved by messages passed within and between factor graph fragments.

The factor graph for model (2.6) that represents the factorization in (3.1) is presented in Figure 1. Each probability density specification in (2.6) is represented by a square node, called a factor, and each of the parameters are represented by circular nodes, called stochastic nodes. The q-density functions that minimize the Kullback-Liebler divergence of the left-hand side of (3.1) from its right-hand side are referred to as optimal q-density functions.

Our presentation of the VMP construction will focus on computing the optimal q-density functions for $\boldsymbol{\nu}$ and $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$. As explained in Minka (2005), the q-density function for $\boldsymbol{\nu}$ and $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$ can be expressed as

$$\begin{aligned} q(\boldsymbol{\nu}) &\propto m_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}}(\boldsymbol{\nu})\; m_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\boldsymbol{\nu}}(\boldsymbol{\nu}) \\ q(\boldsymbol{\zeta}_i) &\propto m_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)\to\boldsymbol{\zeta}_i}(\boldsymbol{\zeta}_i)\; m_{p(\boldsymbol{\zeta}_i)\to\boldsymbol{\zeta}_i}(\boldsymbol{\zeta}_i), \quad i = 1, \ldots, n. \end{aligned} \tag{3.2}$$

Each message has the generic representation $m_{f\to\boldsymbol{\theta}}(\boldsymbol{\theta})$, where f represents an arbitrary factor and $\boldsymbol{\theta}$ represents an arbitrary stochastic node. The arrow in the subscript indicates the direction of the message. Each message is simply a function of the stochastic node that it is sent to or passed from, and their form is described in Minka (2005) and Section 2.5 of Wand (2017).

A key step in deriving and implementing VMP algorithms is the representation of probability density functions in exponential family form: $p(\boldsymbol{x}) \propto \exp\{\boldsymbol{T}(x)^\intercal \boldsymbol{\eta}\}$, where $\boldsymbol{T}(x)$ is a vector of sufficient statistics that identify the distributional family, and $\boldsymbol{\eta}$ is the natural parameter vector; the messages in (3.2) are typically in the exponential family of density functions. Wand (2017) explains how natural parameter vectors play a

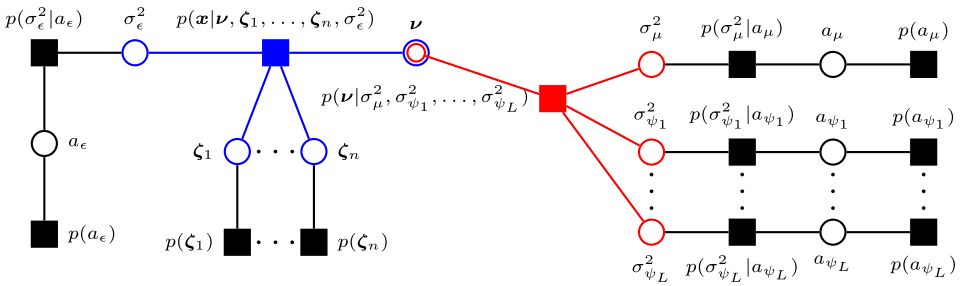

Figure 1: The factor graph for the Bayesian FPCA model in (2.6).

central role in the messages that are passed within and between factor graph fragments. In particular, the natural parameter vectors for the optimal q-density functions in (3.2) take the form

$$\begin{aligned}\boldsymbol{\eta}_{q(\boldsymbol{\nu})} &= \boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\nu}} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\to\boldsymbol{\nu}} \\ \boldsymbol{\eta}_{q(\boldsymbol{\zeta}_i)} &= \boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\zeta}_i} + \boldsymbol{\eta}_{p(\boldsymbol{\zeta}_i)\to\boldsymbol{\zeta}_i}, \quad i = 1,\ldots,n.\end{aligned} \tag{3.3}$$

We outline the exponential family form of the normal and inverse-χ^2 density functions in Appendix B.

We introduce two new fragments that are required for variational inference via VMP for the FPCA model. These are the *functional principal component Gaussian likelihood fragment* (blue in Figure 1) and the *multiple Gaussian penalization fragment* (red in Figure 1). The fragments for $p(\boldsymbol{\zeta}_1),\ldots,p(\boldsymbol{\zeta}_n)$ are *Gaussian prior fragments* (Wand, 2017, Section 4.1.1); the fragments for $p(\sigma_\epsilon^2|a_\epsilon)$, $p(\sigma_\mu^2|a_\mu)$ and $p(\sigma_{\psi_1}^2|a_{\psi_1}),\ldots,p(\sigma_{\psi_L}^2|a_{\psi_L})$ are univariate versions of the *iterated inverse G-Wishart fragment* (Maestrini and Wand, 2021, Algorithm 2); and $p(a_\epsilon)$, $p(a_\mu)$ and $p(a_{\psi_1}),\ldots,p(a_{\psi_L})$ are univariate versions of the *inverse G-Wishart prior fragment* (Maestrini and Wand, 2021, Algorithm 1).

Convergence of the natural parameter vector updates is handled by the notion of minimal Kullback-Leibler divergence. Let $\boldsymbol{\theta}$ represent all the parameters in (2.6). We explain in Appendix E that minimizing the Kullback-Leibler divergence

$$\int q(\boldsymbol{\theta})\log\left\{\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x})}\right\}d\boldsymbol{\theta}$$

is equivalent to maximizing

$$\log \underline{p}(\boldsymbol{x};q) = \int \log\left\{\frac{p(\boldsymbol{x},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\}q(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{3.4}$$

which is a lower-bound on the marginal log-likelihood. The convergence of (3.4) is monitored via coordinate ascent as in Algorithm 1 of Ormerod and Wand (2010).

3.2 Functional Principal Component Gaussian Likelihood Fragment

The message from $p(\boldsymbol{x}|\boldsymbol{\nu}, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n, \sigma_\epsilon^2)$ to $\boldsymbol{\nu}$ can be shown to be proportional to a multivariate normal density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\nu}} \longleftarrow \begin{bmatrix} \mathbb{E}_q(1/\sigma_\epsilon^2) \sum_{i=1}^{n} \left\{ \mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i)^\intercal \otimes \boldsymbol{C}_i \right\}^\intercal \boldsymbol{x}_i \\ -\frac{1}{2}\, \mathbb{E}_q(1/\sigma_\epsilon^2) \sum_{i=1}^{n} \operatorname{vec} \left\{ \mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i \widetilde{\boldsymbol{\zeta}}_i^\intercal) \otimes (\boldsymbol{C}_i^\intercal \boldsymbol{C}_i) \right\} \end{bmatrix}, \tag{3.5}$$

where $\widetilde{\boldsymbol{\zeta}}_i \equiv (1, \boldsymbol{\zeta}_i^\intercal)^\intercal,\ i = 1, \ldots, n$.

For each $i = 1, \ldots, n$, the message from $p(\boldsymbol{x}|\boldsymbol{\nu}, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n, \sigma_\epsilon^2)$ to $\boldsymbol{\zeta}_i$ is proportional to a multivariate normal density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\zeta}_i} \longleftarrow \begin{bmatrix} \mathbb{E}_q(1/\sigma_\epsilon^2) \left\{ \mathbb{E}_q(\boldsymbol{V}_\psi)^\intercal \boldsymbol{C}_i^\intercal \boldsymbol{x}_i - \mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i}) \right\} \\ -\frac{1}{2}\, \mathbb{E}_q(1/\sigma_\epsilon^2) \boldsymbol{D}_L^\intercal \operatorname{vec}\{ \mathbb{E}_q(\boldsymbol{H}_{\psi,i}) \} \end{bmatrix}, \tag{3.6}$$

where $\boldsymbol{V}_\psi \equiv [\ \boldsymbol{\nu}_{\psi_1}\ \ \ldots\ \ \boldsymbol{\nu}_{\psi_L}\ ]$, $\boldsymbol{h}_{\mu\psi,i} \equiv \boldsymbol{V}_\psi^\intercal \boldsymbol{C}_i^\intercal \boldsymbol{C}_i \boldsymbol{\nu}_\mu$ and $\boldsymbol{H}_{\psi,i} \equiv \boldsymbol{V}_\psi^\intercal \boldsymbol{C}_i^\intercal \boldsymbol{C}_i \boldsymbol{V}_\psi$.

The message from $p(\boldsymbol{x}|\boldsymbol{\nu}, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n, \sigma_\epsilon^2)$ to σ_ϵ^2 is an inverse-χ^2 density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\sigma_\epsilon^2} \longleftarrow \begin{bmatrix} -\frac{1}{2} \sum_{i=1}^{n} n_i \\ -\frac{1}{2} \sum_{i=1}^{n} \mathbb{E}_q \left\{ \left(\boldsymbol{x}_i - \boldsymbol{C}_i \boldsymbol{V} \widetilde{\boldsymbol{\zeta}}_i \right)^\intercal \left(\boldsymbol{x}_i - \boldsymbol{C}_i \boldsymbol{V} \widetilde{\boldsymbol{\zeta}}_i \right) \right\} \end{bmatrix}, \tag{3.7}$$

where $\boldsymbol{V} \equiv [\ \boldsymbol{\nu}_\mu\ \ \boldsymbol{\nu}_{\psi_1}\ \ \ldots\ \ \boldsymbol{\nu}_{\psi_L}\ ]$.

Pseudocode for the functional principal component Gaussian likelihood fragment is presented in Algorithm 1. A derivation of all the relevant expectations and natural parameter vector updates is provided in Appendix C.1.

Algorithm 1 Pseudocode for the functional principal component Gaussian likelihood fragment.

Inputs: $\boldsymbol{\eta}_{q(\boldsymbol{\nu})}$, $\{\boldsymbol{\eta}_{q(\boldsymbol{\zeta}_i)} : i = 1, \ldots, n\}$, $\boldsymbol{\eta}_{q(\sigma_\epsilon^2)}$
Updates:
Update posterior expectations. ▷ see Appendix C.1
Update $\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\nu}}$ ▷ equation (3.5)
for $i = 1, \ldots, n$ **do**
    Update $\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\zeta}_i}$ ▷ equation (3.6)
Update $\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\sigma_\epsilon^2}$ ▷ equation (3.7)
Outputs: $\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\nu}}$, $\{\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\zeta}_i} : i = 1, \ldots, n\}$
$\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\sigma_\epsilon^2}$

3.3 Multiple Gaussian Penalization Fragment

The message passed from $p(\boldsymbol{\nu}|\sigma^2_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L})$ to $\boldsymbol{\nu}$ can be shown to be a multivariate normal density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\boldsymbol{\nu}} \longleftarrow \begin{bmatrix} \mathbf{0}_d \\ -\frac{1}{2}\,\mathrm{vec}\left\{\mathbb{E}_q(\boldsymbol{\Sigma}_\nu^{-1})\right\} \end{bmatrix}, \tag{3.8}$$

where $\boldsymbol{\Sigma}_\nu \equiv \mathrm{blockdiag}(\boldsymbol{\Sigma}_\mu, \boldsymbol{\Sigma}_{\psi_1}, \ldots, \boldsymbol{\Sigma}_{\psi_L})$.

The message from $p(\boldsymbol{\nu}|\sigma^2_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L})$ to σ^2_μ is an inverse-χ^2 density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_\mu} \longleftarrow \begin{bmatrix} -\frac{K}{2} \\ -\frac{1}{2}\,\mathbb{E}_q(\boldsymbol{u}_\mu^\intercal \boldsymbol{u}_\mu) \end{bmatrix}. \tag{3.9}$$

Similarly, the message passed from $p(\boldsymbol{\nu}|\sigma^2_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L})$ to $\sigma^2_{\psi_l}$, $l = 1, \ldots, L$, is an inverse-χ^2 density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\psi_l}} \longleftarrow \begin{bmatrix} -\frac{K}{2} \\ -\frac{1}{2}\,\mathbb{E}_q(\boldsymbol{u}_{\psi_l}^\intercal \boldsymbol{u}_{\psi_l}) \end{bmatrix}. \tag{3.10}$$

Pseudocode for the multiple Gaussian penalization fragment is presented in Algorithm 2. A derivation of all the relevant expectations and natural parameter vector updates is provided in Appendix C.2.

Algorithm 2 Pseudocode for the multiple Gaussian penalization fragment.

Inputs: $\boldsymbol{\eta}_{q(\boldsymbol{\nu})}$, $\boldsymbol{\eta}_{q(\sigma^2_\mu)}$, $\{\boldsymbol{\eta}_{q(\sigma^2_{\psi_l})} : l = 1, \ldots, L\}$

Updates:

Update posterior expectations. ▷ see Appendix C.2

Update $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\boldsymbol{\nu}}$ ▷ equation (3.8)

Update $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_\mu}$ ▷ equation (3.9)

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

    Update $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\psi_l}}$ ▷ equation (3.10)

Outputs: $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\boldsymbol{\nu}}$, $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_\mu}$, $\{\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\psi_l}} : l = 1, \ldots, L\}$

Note that the multiple Gaussian penalization fragment is not a fragment designed specifically for orthogonal FPCA because it does not account for an orthogonal set of eigenfunctions. Indeed, it can be applied to any statistical model that specifies an independent penalization structure over multiple vectors of spline coefficients.

4 Multilevel Extensions

Di et al. (2009) introduced multilevel FPCA (MlFPCA) for clustered or multilevel functional data. The multilevel extension of (2.2), which is a simplified version the

MlFPCA model in Di et al. (2009), is

$$\widehat{x}_{ij}(t) = \mu(t) + \sum_{l=1}^{L_1} \zeta_{il}^{(1)} \psi_l^{(1)}(t) + \sum_{l=1}^{L_2} \zeta_{ijl}^{(2)} \psi_l^{(2)}(t), \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, n. \quad (4.1)$$

The first level eigenfunctions $\{\psi_l^{(1)}\}_{l=1,\ldots,L_1}$ account for first level specific shifts from the mean function $\mu(t)$, and the second level eigenfunctions $\{\psi_l^{(2)}\}_{l=1,\ldots,L_2}$ account for second level specific shifts from the subject-specific mean function. Note that the level one and level two eigenfunctions form an orthonormal basis, but are not required to be mutually orthogonal. With the assumption of Gaussian residuals, the first level scores $\zeta_{il}^{(1)}$ are uncorrelated zero-mean Gaussian random variables, and likewise for the second level scores $\zeta_{ijl}^{(2)}$. Additionally, first and second level scores are assumed to be uncorrelated. A similar MlFPCA decomposition was used in Goldsmith et al. (2015) for generalized multilevel function-on-scalar regression.

In constructing the Bayesian model, we note that the multivariate extension of (2.4) is

$$\boldsymbol{x}_{ij} = \boldsymbol{\mu}_{ij} + \sum_{l=1}^{L_1} \zeta_{il}^{(1)} \boldsymbol{\psi}_{ijl}^{(1)} + \sum_{l=1}^{L_2} \zeta_{ijl}^{(2)} \boldsymbol{\psi}_{ijl}^{(2)} + \boldsymbol{\epsilon}_{ij}, \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{x}_{ij}, \boldsymbol{\mu}_{ij}, \{\boldsymbol{\psi}_{ijl}^{(1)}\}_{l=1,\ldots,L_1}$ and $\{\boldsymbol{\psi}_{ijl}^{(2)}\}_{l=1,\ldots,L_2}$ are $n_{ij} \times 1$ vectors defined over the design points $t_{ij1}, \ldots, t_{ijn_{ij}}$ for the ijth observation. For nonparametric fitting of the nonlinear curves in the model, the design matrix $\boldsymbol{C}_{ij}$ has an identical structure to (2.5), but evaluated of the design points $t_{ij1}, \ldots, t_{ijn_{ij}}$. Vectors of spline coefficients are defined as $\boldsymbol{\nu}_\mu, \{\boldsymbol{\nu}_{\psi_l}^{(1)}\}_{l=1,\ldots,L_1}$ and $\{\boldsymbol{\nu}_{\psi_l}^{(2)}\}_{l=1,\ldots,L_2}$ for the mean function, the first level eigenfunctions, and the second level eigenfunctions, respectively. Vectors of scores are defined such that $\boldsymbol{\zeta}_i^{(1)} \equiv (\zeta_{i1}^{(1)}, \ldots, \zeta_{iL_1}^{(1)})^\intercal$ and $\boldsymbol{\zeta}_{ij}^{(2)} \equiv (\zeta_{ij1}^{(2)}, \ldots, \zeta_{ijL_2}^{(1)})^\intercal$. In addition, we will introduce a slight abuse of notation by setting $L \equiv L_1 + L_2$, $\boldsymbol{\nu} \equiv (\boldsymbol{\nu}_\mu^\intercal, \boldsymbol{\nu}_{\psi_1}^{(1)\intercal}, \ldots, \boldsymbol{\nu}_{\psi_{L_1}}^{(1)\intercal}, \boldsymbol{\nu}_{\psi_1}^{(2)\intercal}, \ldots, \boldsymbol{\nu}_{\psi_{L_2}}^{(2)\intercal})^\intercal$ and $\boldsymbol{\zeta}_i \equiv (\boldsymbol{\zeta}_i^{(1)\intercal}, \boldsymbol{\zeta}_{i1}^{(2)\intercal}, \ldots, \boldsymbol{\zeta}_{im_i}^{(2)\intercal})^\intercal$ for $i = 1, \ldots, n$ in the multilevel setting. However, the precise definition of L, $\boldsymbol{\nu}$ and $\boldsymbol{\zeta}_i$ should be apparent from the specific model (standard or multilevel) that we are analysing.

The Bayesian MlFPCA model is:

$$\boldsymbol{x}_{ij} | \boldsymbol{\nu}, \boldsymbol{\zeta}_i^{(1)}, \boldsymbol{\zeta}_{ij}^{(2)}, \sigma_\epsilon^2 \overset{\text{ind.}}{\sim} \mathrm{N}\left\{ \boldsymbol{C}_{ij} \left(\boldsymbol{\nu}_\mu + \sum_{l=1}^{L_1} \zeta_{il}^{(1)} \boldsymbol{\nu}_{\psi_l}^{(1)} + \sum_{l=1}^{L_2} \zeta_{ijl}^{(2)} \boldsymbol{\nu}_{\psi_l}^{(2)} \right), \sigma_\epsilon^2 \boldsymbol{I}_{n_{ij}} \right\},$$

$$\boldsymbol{\zeta}_i^{(1)} \overset{\text{ind.}}{\sim} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{\zeta}_{ij}^{(2)} \overset{\text{ind.}}{\sim} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}), \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, n,$$

$$\begin{bmatrix} \boldsymbol{\nu}_\mu \\ \boldsymbol{\nu}_{\psi_l}^{(1)} \\ \boldsymbol{\nu}_{\psi_k}^{(2)} \end{bmatrix} \Bigg| \; \sigma_\mu^2, \sigma_{\psi_l}^{(1)2}, \sigma_{\psi_k}^{(2)2} \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_\mu & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_{\psi_l}^{(1)} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\psi_k}^{(2)} \end{bmatrix} \right),$$

$$\sigma_{\psi_l}^{(1)2} | a_{\psi_l}^{(1)} \overset{\text{ind.}}{\sim} \text{Inverse} - \chi^2(1, 1/a_{\psi_l}^{(1)}), \quad a_{\psi_l}^{(1)} \overset{\text{ind.}}{\sim} \text{Inverse} - \chi^2(1, 1/A^2), \quad l = 1, \ldots, L_1,$$

$$\begin{aligned}\sigma^{(2)2}_{\psi_k}|a^{(2)}_{\psi_k} &\overset{\text{ind.}}{\sim} \text{Inverse}-\chi^2(1,1/a^{(2)}_{\psi_k}), \quad a^{(2)}_{\psi_k} \overset{\text{ind.}}{\sim} \text{Inverse}-\chi^2(1,1/A^2), \quad k=1,\ldots,L_2,\\ \sigma^2_\mu|a_\mu &\sim \text{Inverse}-\chi^2(1,1/a_\mu), \quad a_\mu \sim \text{Inverse}-\chi^2(1,1/A^2),\\ \sigma^2_\epsilon|a_\epsilon &\sim \text{Inverse}-\chi^2(1,1/a_\epsilon), \quad a_\epsilon \sim \text{Inverse}-\chi^2(1,1/A^2).\end{aligned} \tag{4.2}$$

4.1 VMP for MIFPCA

The mean field restriction that we set for model (4.2), after applying induced factorizations similar to those used to derive (3.1), is:

$$\begin{aligned}&q(\boldsymbol{\nu},\{\boldsymbol{\zeta}_i\}_{i=1,\ldots,n},\sigma^2_\epsilon,a_\epsilon,\sigma^2_\mu,a_\mu,\{\sigma^{(1)2}_{\psi_l},a^{(1)}_{\psi_l}\}_{l=1,\ldots,L_1},\{\sigma^{(2)2}_{\psi_l},a^{(2)}_{\psi_l}\}_{l=1,\ldots,L_2})\\ &= \left\{\prod_{i=1}^n q(\boldsymbol{\zeta}_i)\right\} q(\boldsymbol{\nu})q(\sigma^2_\epsilon)q(a_\epsilon)q(\sigma^2_\mu)q(a_\mu)\left\{\prod_{l=1}^{L_1} q(\sigma^{(1)2}_{\psi_l})q(a^{(1)}_{\psi_l})\right\}\left\{\prod_{l=1}^{L_2} q(\sigma^{(2)2}_{\psi_l})q(a^{(2)}_{\psi_l})\right\}.\end{aligned} \tag{4.3}$$

The factor graph for model (4.2) that represents the factorization in (4.3) is presented in Figure 2.

In extending the Bayesian FPCA model to its multilevel form, we can see the advantage of using a VMP approach to variational Bayesian inference. First note the fragments that have been derived in previous publications: the fragments for $p(\sigma^2_\epsilon|a_\epsilon)$, $p(\sigma^{(1)2}_{\psi_1}|a^{(1)}_{\psi_1}),\ldots,p(\sigma^{(1)2}_{\psi_{L_1}}|a^{(1)}_{\psi_{L_1}})$ and $p(\sigma^{(2)2}_{\psi_1}|a^{(2)}_{\psi_1}),\ldots,p(\sigma^{(2)2}_{\psi_{L_2}}|a^{(2)}_{\psi_{L_2}})$ are univariate versions of the iterated inverse G-Wishart fragment; and the fragments for $p(a_\epsilon)$, $p(a^{(1)}_{\psi_1})$, $\ldots,p(a^{(1)}_{\psi_{L_1}})$ and $p(a^{(1)}_{\psi_1}),\ldots,p(a^{(1)}_{\psi_{L_1}})$ are univariate versions of the inverse G-Wishart prior fragment. Next, we focus on the fragment colored in red in Figure 2, which computes the natural parameter vector updates for messages passed from $p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^{(1)2}_{\psi_1},\ldots,$ $\sigma^{(1)2}_{\psi_{L_1}},\sigma^{(2)2}_{\psi_1},\ldots,\sigma^{(2)2}_{\psi_{L_2}})$. Notice that the form of this probabilistic specification in (4.2) involves a set of L_1+L_2+1 vectors with independent Gaussian penalization specifications. Following on from the discussion at the end of Section 3.3, this probabilistic specification is identical to the multiple Gaussian penalization specification in (2.6). Therefore, the updates for the multiple Gaussian penalization fragment in Algorithm 2 can be recycled into the multilevel model to compute the natural parameter vector updates for the red fragment in Figure 2. Next, in extending the FPCA model to a multilevel model through a VMP scheme, we must derive the updates for the fragment colored in blue in Figure 2, which we name the *multilevel functional principal component Gaussian likelihood fragment*. In addition, we must also modify the update for the multilevel Gaussian prior fragments represented by $p(\boldsymbol{\zeta}_1),\ldots,p(\boldsymbol{\zeta}_n)$. However, these prior updates are simple in form and only need to be computed once when running VMP algorithms (Wand, 2017).

Here, we see that addressing FPCA with a VMP approach naturally permits derivational and computational savings. If we were to address this model with traditional MFVB updates, we would have to re-derive and re-code the parameters of the optimal

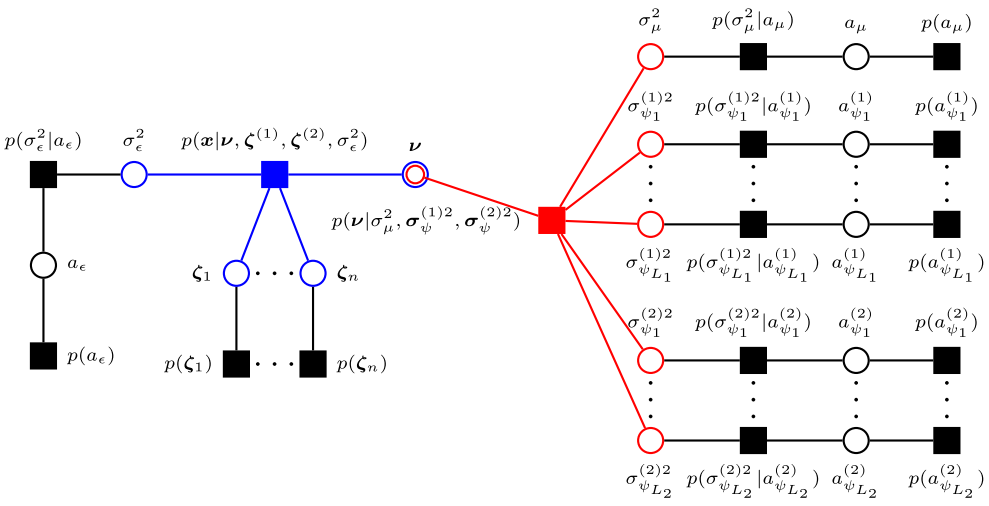

Figure 2: The factor graph for the Bayesian MlFPCA model in (4.2). To avoid notational clutter within the factor graph, we have introduced some notation: we set $\boldsymbol{\sigma}_\psi^{(1)2} = \{\sigma_{\psi_l}^{(1)2}\}_{l=1,\ldots,L_1}$ and $\boldsymbol{\sigma}_\psi^{(1)2} = \{\sigma_{\psi_l}^{(2)2}\}_{l=1,\ldots,L_2}$.

posterior density functions of most of the model parameters. Under a VMP scheme, however, we only need to derive the updates for a new fragment, the multilevel functional principal component Gaussian likelihood fragment, and modify the update for the Gaussian prior fragment.

4.2 Multilevel Functional Principal Component Gaussian Likelihood Fragment

The message from $p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)$ to $\boldsymbol{\nu}$ can be shown to be proportional to a multivariate normal density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\nu}} \longleftarrow \begin{bmatrix} \mathbb{E}_q(1/\sigma_\epsilon^2)\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_{ij})^\intercal \otimes \boldsymbol{C}_{ij}\right\}^\intercal \boldsymbol{x}_{ij} \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma_\epsilon^2)\sum_{i=1}^{n}\sum_{j=1}^{m}\operatorname{vec}\left\{\mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_{ij}\widetilde{\boldsymbol{\zeta}}_{ij}^\intercal)\otimes(\boldsymbol{C}_{ij}^\intercal\boldsymbol{C}_{ij})\right\} \end{bmatrix}, \quad (4.4)$$

where $\widetilde{\boldsymbol{\zeta}}_i \equiv (1, \boldsymbol{\zeta}_i^{(1)\intercal}, \boldsymbol{\zeta}_{ij}^{(2)\intercal})^\intercal$, for $j = 1,\ldots,m_i$ and $i = 1,\ldots,n$.

In the multilevel setting, we set $\boldsymbol{V}_\psi^{(r)} \equiv [\ \boldsymbol{\nu}_{\psi_1}^{(r)} \ \cdots \ \boldsymbol{\nu}_{\psi_{L_r}}^{(r)}\ ]$, for $r = 1, 2$, $\boldsymbol{C}_i \equiv \text{stack}(\{\boldsymbol{C}_{ij}\}_{j=1,\ldots,m_i})$, $\boldsymbol{C}_{\psi,i}^{(1)} \equiv \boldsymbol{C}_i\boldsymbol{V}_\psi^{(1)}$, $\boldsymbol{C}_{\psi,i}^{(2)} \equiv \text{blockdiag}(\{\boldsymbol{C}_{ij}\boldsymbol{V}_\psi^{(2)}\}_{j=1,\ldots,m_i})$, $\boldsymbol{C}_{\psi,i} \equiv [\ \boldsymbol{C}_{\psi,i}^{(1)} \ \ \boldsymbol{C}_{\psi,i}^{(2)}\ ]$ and $\boldsymbol{H}_{\psi,i} \equiv \boldsymbol{C}_{\psi,i}^\intercal\boldsymbol{C}_{\psi,i}$. Note that, for $i = 1,\ldots,n$, the full conditional of $\boldsymbol{\zeta}_i$ is a multivariate normal with inverse covariance matrix $\sigma_\epsilon^{-2}\boldsymbol{H}_{\psi,i} + \boldsymbol{I}_{L_1+m_iL_2}$,

which is a two-level sparse matrix (Nolan and Wand, 2020, Definition 1). We provide a brief overview of the two-level sparse matrix problems of Nolan and Wand (2020) in Appendix C.3, which permit streamlined computations of natural parameter vector updates without directly inverting a two-level sparse matrix.

Sections 4.3 and 4.4 of Nolan et al. (2020) provide streamlined computations for VMP fragments that facilitate variational Bayesian inference for two-level linear mixed models. In order to use similar methods for streamlining the messages passed to each $\{\boldsymbol{\zeta}_i\}_{i=1,\ldots,n}$, we must adopt the reduced exponential family forms for multivariate normal messages in Nolan et al. (2020). Under these circumstances, the message from $p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_i,\sigma^2_\epsilon)$ to $\boldsymbol{\zeta}_i$ is

$$m_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_i,\sigma^2_\epsilon)\to\boldsymbol{\zeta}_i}(\boldsymbol{\zeta}_i) = \exp\left\{\begin{bmatrix} \boldsymbol{\zeta}_i^{(1)} \\ \text{vech}(\boldsymbol{\zeta}_i^{(1)}\boldsymbol{\zeta}_i^{(1)\intercal}) \\ \underset{j=1,\ldots,m_i}{\text{stack}}\left\{\begin{bmatrix} \boldsymbol{\zeta}_{ij}^{(2)} \\ \text{vech}(\boldsymbol{\zeta}_{ij}^{(2)}\boldsymbol{\zeta}_{ij}^{(2)\intercal}) \\ \text{vec}(\boldsymbol{\zeta}_i^{(1)}\boldsymbol{\zeta}_{ij}^{(2)\intercal}) \end{bmatrix}\right\} \end{bmatrix}^\intercal \boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_i,\sigma^2_\epsilon)\to\boldsymbol{\zeta}_i}\right\}, \tag{4.5}$$

which is proportional to a multivariate normal density function in reduced exponential form. Next, set $\boldsymbol{C}^{(r)}_{\psi,ij} \equiv \boldsymbol{C}_{ij}\boldsymbol{V}^{(r)}_\psi$, $\boldsymbol{h}^{(r)}_{\mu\psi,ij} \equiv \boldsymbol{V}^{(r)\intercal}_\psi \boldsymbol{C}^\intercal_{ij}\boldsymbol{C}_{ij}\boldsymbol{\nu}_\mu$, $\boldsymbol{H}^{(r,s)}_{\psi,ij} \equiv \boldsymbol{C}^{(r)\intercal}_{\psi,ij}\boldsymbol{C}^{(s)}_{\psi,ij}$, for $r,s = 1,2$. Then, for $i = 1,\ldots,n$, the natural parameter vector update in (4.5) is

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_i,\sigma^2_\epsilon)\to\boldsymbol{\zeta}_i} \longleftarrow \begin{bmatrix} \mathbb{E}_q(1/\sigma^2_\epsilon)\sum_{j=1}^m\{\mathbb{E}_q(\boldsymbol{C}^{(1)}_{\psi,ij})^\intercal\boldsymbol{x}_{ij} - \mathbb{E}_q(\boldsymbol{h}^{(1)}_{\mu\psi,ij})\} \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2_\epsilon)\boldsymbol{D}^\intercal_{L_1}\sum_{j=1}^m \text{vech}\{\mathbb{E}_q(\boldsymbol{H}^{(1,1)}_{\psi,ij})\} \\ \underset{j=1,\ldots,m_i}{\text{stack}}\left(\begin{bmatrix} \mathbb{E}_q(1/\sigma^2_\epsilon)\{\mathbb{E}_q(\boldsymbol{C}^{(2)}_{\psi,ij})^\intercal\boldsymbol{x}_{ij} - \mathbb{E}_q(\boldsymbol{h}^{(2)}_{\mu\psi,ij})\} \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2_\epsilon)\boldsymbol{D}^\intercal_{L_2}\,\text{vec}\{\mathbb{E}_q(\boldsymbol{H}^{(2,2)}_{\psi,ij})\} \\ -\,\mathbb{E}_q(1/\sigma^2_\epsilon)\,\mathbb{E}_q\,\text{vec}\{\mathbb{E}_q(\boldsymbol{H}^{(1,2)}_{\psi,ij})\} \end{bmatrix}\right) \end{bmatrix}. \tag{4.6}$$

The message from $p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\epsilon)$ to σ^2_ϵ is an inverse-χ^2 density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\epsilon)\to\sigma^2_\epsilon} \longleftarrow \begin{bmatrix} -\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^{m_i} n_{ij} \\ -\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^{m_i}\mathbb{E}_q\left\{\left|\left|\boldsymbol{x}_{ij} - \boldsymbol{C}_{ij}\left(\boldsymbol{\nu}_\mu - \boldsymbol{V}^{(1)}_\psi\boldsymbol{\zeta}^{(1)}_i - \boldsymbol{V}^{(2)}_\psi\boldsymbol{\zeta}^{(2)}_{ij}\right)\right|\right|^2\right\} \end{bmatrix}. \tag{4.7}$$

Pseudocode for the multilevel functional principal component Gaussian likelihood fragment is presented in Algorithm 3. A derivation of all the relevant expectations and natural parameter vector updates is provided in Appendix C.4.

Algorithm 3 Pseudocode for the multilevel functional principal component Gaussian likelihood fragment.

Inputs: $\boldsymbol{\eta}_{q(\boldsymbol{\nu})}$, $\{\boldsymbol{\eta}_{q(\boldsymbol{\zeta}_i)} : i = 1, \ldots, n\}$, $\boldsymbol{\eta}_{q(\sigma_\epsilon^2)}$
Updates:
Update posterior expectations. ▷ see Appendix C.4
Update $\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\nu}}$ ▷ equation (4.4)
for $i = 1, \ldots, n$ **do**
    Update $\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\zeta}_i}$ ▷ equation (4.6)
Update $\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\sigma_\epsilon^2}$ ▷ equation (4.7)
Outputs: $\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\nu}}$, $\{\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\boldsymbol{\zeta}_i} : i = 1, \ldots, n\}$
$\{\boldsymbol{\eta}_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\sigma_\epsilon^2}$, $G_{p(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma_\epsilon^2)\to\sigma_\epsilon^2}\}$

4.3 Multilevel Gaussian Prior Fragment

In order to preserve conjugate messages passed to each $\boldsymbol{\zeta}_i$, we need to adjust the computations for the Gaussian prior fragment. More specifically, we simply need to transform the messages into the form of (4.5). Here, the natural parameter vector update is

$$\boldsymbol{\eta}_{p(\boldsymbol{\zeta}_i)\to\boldsymbol{\zeta}_i} \longleftarrow \begin{bmatrix} \mathbf{0}_{L_1} \\ -\frac{1}{2}\boldsymbol{D}_{L_1}^\intercal \operatorname{vech}(\boldsymbol{I}_{L_1}) \\ \mathbf{1}_{m_i} \otimes \begin{bmatrix} \mathbf{0}_{L_2} \\ -\frac{1}{2}\boldsymbol{D}_{L_2}^\intercal \otimes \operatorname{vech}(\boldsymbol{I}_{L_2}) \end{bmatrix} \\ \mathbf{0}_{L_1 L_2} \end{bmatrix}. \tag{4.8}$$

5 Post-VMP Steps

The FPCA model for curve estimation (2.2), which has its genesis in the Karhunen-Loève decomposition (2.1), relies on orthogonal functional principal component eigenfunctions and independent vectors of scores with uncorrelated entries. However, the variational Bayesian FPCA resulting from a VMP treatment does not enforce any orthogonality restrictions on the resulting eigenfunctions. Although prediction of the response curves is still valid without these constraints, interpretation of the analysis is more straightforward with orthogonal eigenfunctions. In the following sections, we outline a sequence of post-VMP steps that aid inference and interpretability for variational Bayes-based FPCA.

5.1 Establishing the Optimal Posterior Density Functions

We are primarily concerned with the optimal posterior density functions for the vector of spline coefficients for the mean function and eigenfunctions $\boldsymbol{\nu}$ and the vectors of principal component scores $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$. Upon convergence of the VMP algorithm, the natural parameter vectors for these optimal posterior density functions can be computed via (3.3). The optimal posterior density for each of these parameters is a Gaussian density function, where the mean vector $\mathbb{E}_q(\boldsymbol{\nu})$ and covariance matrix $\mathbb{C}\text{ov}_q(\boldsymbol{\nu})$ of $q(\boldsymbol{\nu})$

can be computed from (B.2), and the corresponding parameters $\mathbb{E}_q(\boldsymbol{\zeta}_i)$ and $\mathbb{C}\text{ov}_q(\boldsymbol{\zeta}_i)$ of $q(\boldsymbol{\zeta}_i)$, $i = 1, \ldots, n$, can be computed from (B.3). Note that we partition $\mathbb{E}_q(\boldsymbol{\nu})$ as $\mathbb{E}_q(\boldsymbol{\nu}) = \{\mathbb{E}_q(\boldsymbol{\nu}_\mu)^\intercal, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_1})^\intercal, \ldots, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_L})^\intercal\}^\intercal$.

5.2 Posterior Estimation of the Karhunen-Loève Decomposition

In this section, we outline a sequence of steps to establish orthogonal functional principal component eigenfunctions and uncorrelated scores. Note that we will treat the estimated functional principal component eigenfunctions as fixed curves that are estimated from the posterior mean of the spline coefficients $\mathbb{E}_q(\boldsymbol{\nu})$. As a consequence, the pointwise posterior variance in the response curve estimates result from the variance in the principal component scores alone. This treatment is in line with standard approaches to FPCA, where the randomness in the model is generated by the scores (e.g. Yao et al., 2005; Benko et al., 2009).

Now, we outline the steps to construct orthogonal functional principal component eigenfunctions and uncorrelated scores. The existence and uniqueness of the eigenfunctions, up to a change of sign, are justified by Theorem 2.1. First, set up an equidistant grid of design points $\boldsymbol{t}_g = (t_{g1}, \ldots, t_{gn_g})^\intercal$, where $t_{g1} = 0$, $t_{gn_g} = 1$ and n_g is the length of the grid. Then define $\boldsymbol{C}_g$ in an analogous fashion to (2.5): $\boldsymbol{C}_g \equiv [\ \mathbf{1}_{n_g} \quad \boldsymbol{t}_g \quad z_1(\boldsymbol{t}_g) \quad \cdots \quad z_K(\boldsymbol{t}_g)\ ]$, where $\mathbf{1}_{n_g}$ is an $n_g \times 1$ vector of ones. The posterior estimate of the mean function is

$$\widehat{\mu}(\boldsymbol{t}_g) \equiv \mathbb{E}_q\{\mu(\boldsymbol{t}_g)\} = \boldsymbol{C}_g\, \mathbb{E}_q(\boldsymbol{\nu}_\mu). \tag{5.1}$$

Next, the variational Bayes estimates of the functional principal components eigenfunctions are $\mathbb{E}_q\{\psi_l(\boldsymbol{t}_g)\} = \boldsymbol{C}_g\, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})$, $l = 1, \ldots, L$. Then define the matrix $\boldsymbol{\Psi}$ such that $\boldsymbol{\Psi} \equiv [\ \mathbb{E}_q\{\psi_1(t_g)\} \quad \cdots \quad \mathbb{E}_q\{\psi_L(t_g)\}\ ]$. Establish the singular value decomposition of $\boldsymbol{\Psi}$ such that $\boldsymbol{\Psi} = \boldsymbol{U}_\psi \boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal$, where $\boldsymbol{U}_\psi$ is an $n_g \times L$ matrix consisting of the first L left singular vectors of $\boldsymbol{\Psi}$, $\boldsymbol{V}_\psi$ is an $L \times L$ matrix consisting of the right singular vectors of $\boldsymbol{\Psi}$, and $\boldsymbol{D}_\psi$ is an $L \times L$ diagonal matrix consisting of the singular values of $\boldsymbol{\Psi}$.

Now, define $\boldsymbol{\Xi} \equiv [\ \mathbb{E}_q(\boldsymbol{\zeta}_1) \quad \cdots \quad \mathbb{E}_q(\boldsymbol{\zeta}_n)\ ]^\intercal$. Then set $\boldsymbol{C}_\zeta$ to be the $L \times L$ sample covariance matrix of the column vectors of $\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal \boldsymbol{\Xi}^\intercal$ and establish its spectral decomposition $\boldsymbol{C}_\zeta = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\intercal$, where $\boldsymbol{\Lambda}$ is a diagonal matrix consisting of the eigenvalues of $\boldsymbol{C}_\zeta$ in descending order along its main diagonal and $\boldsymbol{Q}$ is the orthogonal matrix consisting of the corresponding eigenvectors of $\boldsymbol{C}_\zeta$ along its columns.

Finally, define the matrices

$$\dot{\boldsymbol{\Psi}} \equiv \boldsymbol{U}_\psi \boldsymbol{Q} \boldsymbol{\Lambda}^{1/2} \quad \text{and} \quad \dot{\boldsymbol{\Xi}} \equiv \boldsymbol{\Xi} \boldsymbol{V}_\psi \boldsymbol{D}_\psi \boldsymbol{Q} \boldsymbol{\Lambda}^{-1/2}. \tag{5.2}$$

Notice that $\dot{\boldsymbol{\Psi}}$ is an $n_g \times L$ matrix and $\dot{\boldsymbol{\Xi}}$ is an $n \times L$ matrix. Next, partition these matrices such that the lth column of $\dot{\boldsymbol{\Psi}}$ is $\dot{\psi}_l(\boldsymbol{t}_g)$ and the ith row of $\dot{\boldsymbol{\Xi}}$ is $(\dot{\zeta}_{i1}, \ldots, \dot{\zeta}_{iL})$. The columns of $\dot{\boldsymbol{\Psi}}$ are orthonormal vectors, but we require continuous curves that are orthonormal in $L^2([0,1])$. We can adjust this by finding an approximation of $||\dot{\psi}_l||$,

$l = 1, \ldots, L$, through numerical integration. This allows us to establish estimates of the orthonormal functions $\psi_1, \ldots, \psi_L$ in (2.2) over the vector $\boldsymbol{t}_g$ with

$$\widehat{\psi}_l(\boldsymbol{t}_g) \equiv \frac{\dot{\psi}_l(\boldsymbol{t}_g)}{||\dot{\psi}_l||}, \quad l = 1, \ldots, L, \tag{5.3}$$

as well as estimates of the scores with $\widehat{\zeta}_{il} \equiv ||\dot{\psi}_l||\dot{\zeta}_{il}$. Lemma 5.1 outlines the construction of posterior curve estimation for the response vectors $x_1(\boldsymbol{t}_g), \ldots, x_n(\boldsymbol{t}_g)$. Proposition 5.1 shows that the form of the predicted response vectors in Lemma 5.1 is a discrete version of the Karhunen-Loève decomposition. Here, we define $\widehat{\boldsymbol{\zeta}}_i \equiv (\widehat{\zeta}_{i1}, \ldots, \widehat{\zeta}_{iL})^\intercal$, $i = 1, \ldots, n$.

Lemma 5.1. *The posterior estimate for the response vector $x_i(\boldsymbol{t}_g)$ is given by*

$$\widehat{x}_i(\boldsymbol{t}_g) = \widehat{\mu}(\boldsymbol{t}_g) + \sum_{l=1}^{L} \widehat{\zeta}_{il}\widehat{\psi}_l(\boldsymbol{t}_g), \quad i = 1, \ldots, n. \tag{5.4}$$

Remark. *The posterior estimates $\widehat{x}_1(\boldsymbol{t}_g), \ldots, \widehat{x}_n(\boldsymbol{t}_g)$ in (5.4) are the same as those prior to the post-processing steps. That is, $\widehat{x}_i(\boldsymbol{t}_g) = \boldsymbol{C}_g\, \mathbb{E}_q(\boldsymbol{\nu}_\mu) + \sum_{l=1}^{L} \mathbb{E}_q(\zeta_{il})\boldsymbol{C}_g\, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})$. In summary, the post processing steps simply orthogonalize and normalize the eigenfunctions and uncorrelate the scores, but do not affect the fits to the observed data.*

Proposition 5.1. *The vectors $\widehat{\boldsymbol{\zeta}}_1, \ldots, \widehat{\boldsymbol{\zeta}}_N$ are independent with sample covariance matrix* $\text{diag}(||\dot{\psi}_1||^2, \ldots, ||\dot{\psi}_L||^2)$. *Furthermore, the vectors $\widehat{\psi}_1(\boldsymbol{t}_g), \ldots, \widehat{\psi}_L(\boldsymbol{t}_g)$ are eigenvectors of the sample covariance matrix of $\widehat{x}_1(\boldsymbol{t}_g), \ldots, \widehat{x}_n(\boldsymbol{t}_g)$, and $||\dot{\psi}_1||^2, \ldots, ||\dot{\psi}_L||^2$ are the corresponding eigenvalues.*

The proof of Lemma 5.1 is presented in Appendix A.2, and the proof of Proposition 5.1 is presented in Appendix A.3.

These steps can be naturally extended to the multilevel setting by applying them separately to each level. A clear outline is provided in Appendix D.

6 Simulations

We illustrate the use of Algorithms 1, 2 and 3 through a series of simulations of models (2.6) and (4.2). Pseudocode for the VMP algorithm for the standard FPCA model (2.6) and the MlFPCA model (4.2) are provided in Algorithms 1 and 2 of Appendix E.2. The VMP algorithms were determined to have converged once the relative increase in $\log \underline{p}(\boldsymbol{x}; q)$ fell below 10^{-5}. In addition, we have included the results from MCMC treatments of both models for comparison with the VMP-based variational Bayesian inference. MCMC simulations were conducted through Rstan, the R (R Core Team, 2020) interface to the probabilistic programming language Stan (Stan Development Team, 2020). For each simulation, we used Rstan's default no-U-turn sampler with 1000 burn-in samples and 1000 MCMC samples.

6.1 Accuracy Assessment

Simulations of the FPCA model (2.6) were conducted with $n \in \{10, 50, 100, 250, 500\}$. The number of observations n_i for the ith curve was sampled uniformly over $\{20, 21, \ldots, 30\}$, while the time observations for the ith curve $\{t_{i1}, \ldots, t_{in_i}\}$ were sampled uniformly over the interval $(0, 1)$. The residual variance σ_ϵ^2 was set to 1. The mean function was $\mu(t) = 3\sin(\pi t) - 1.5$ and the eigenfunctions were $\psi_1(t) = \sqrt{2}\sin(2\pi t)$, $\psi_2(t) = \sqrt{2}\cos(2\pi t)$, $\psi_3(t) = \sqrt{2}\sin(4\pi t)$ and $\psi_4(t) = \sqrt{2}\cos(4\pi t)$. Each principal component score was simulated according to $\zeta_{il} \overset{\text{ind.}}{\sim} \text{N}(0, 1/l^2)$, $i = 1, \ldots, n$, $l = 1, \ldots, 4$. Hyperparameter specifications were $\sigma_\beta^2 = 10^{10}$ and $A = 10^5$, ensuring arbitrarily uninformative priors on fixed-effects parameters and standard deviation parameters. For each $n \in \{10, 50, 100, 250, 500\}$, we conducted 100 simulations of model (2.6) with the aim of analysing the error of the posterior estimates of the eigenfunctions. The error of each simulation was determined via the integrated squared error:

$$\text{ISE}(\psi_l, \widehat{\psi}_l) = \int_0^1 \left|\psi_l(t) - \widehat{\psi}_l(t)\right|^2 dt, \quad l = 1, \ldots, L. \tag{6.1}$$

For comparison, we also present the analogous accuracy scores for the MCMC algorithms. We conducted a similar series of simulations for model (4.2) with the additional parameters m_i sampled uniformly over $\{10, 11, \ldots, 15\}$, $\psi_1^{(1)}(t) = \sqrt{2}\sin(2\pi t)$, $\psi_2^{(1)}(t) = \sqrt{2}\cos(2\pi t)$, $\psi_3^{(1)}(t) = \sqrt{2}\sin(4\pi t)$, $\psi_1^{(2)}(t) = \sqrt{2}\cos(4\pi t)$, $\psi_2^{(2)}(t) = \sqrt{2}\sin(6\pi t)$, $\psi_3^{(2)}(t) = \sqrt{2}\cos(6\pi t)$ and $\zeta_{il}^{(r)} \overset{\text{ind.}}{\sim} \text{N}(0, 1/l^2)$, $i = 1, \ldots, 50$, $l = 1, 2, 3$, $r = 1, 2$.

Nonparameteric regression with O'Sullivan penalized splines for the nonlinear curves was performed with $K = 12$. Ruppert (2002) sets a simple default value for K as $\min(n_{\text{obs}}/4, 40)$, where n_{obs} is the number of observations. However, one of the general conclusions from this article is that there is a minimum adequate value for K, with regressions exceeding this value giving satisfactory fits because the penalty prevents overfitting. Since there are 20–30 observations per subject in our simulations, we treat n_{obs} as 30 and use the simple default setting of Ruppert (2002) as a proxy for the minimum adequate value for K. By setting $K = 12$, we exceed the minimum setting, ensuring adequate fits.

The box plots for the logarithm of the integrated squared error values for the FPCA model in Figure 3 reflect the overall excellent results of the VMP algorithms. For lower values of n, the log ISE for $\psi_4(t)$ tends to be greater for the VMP simulations than the MCMC simulations. This fall off in accuracy for the VMP simulations is a result of the weak contribution of this eigenfunction to the variability of the dataset, where it accounts for approximately 4.4% of the total variability. Similar results are also observed for the MlFPCA simulations. In addition, we provide a simple illustration of the fits for the FPCA and MlFPCA models in Appendix F.

To assess the accuracy in the estimation of the scores, we used the root mean squared error (RMSE). The results are listed in Table 1 for the FPCA and the MlFPCA models via VMP and MCMC. The RMSE values for the posterior estimates of the scores in the VMP simulations match well with the MCMC simulations.

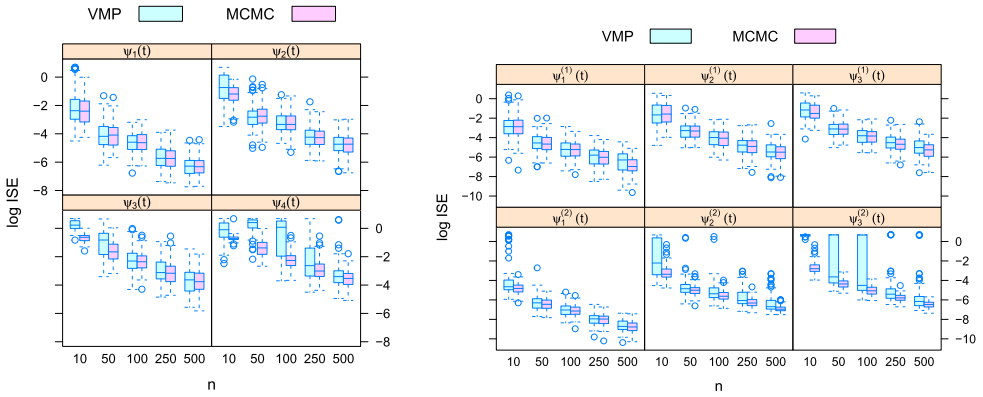

Figure 3: *Left panel*: The results from a simulation study of the Bayesian FPCA model in (2.6). *Right panel*: Analogous results for the Bayesian MlFPCA model in (4.2). The box plots in each panel are a summary of the logarithm of the integrated squared error values in (6.1) for 100 simulations of each of the settings $n \in \{10, 50, 100, 250, 500\}$. We have also included the corresponding accuracy results for the MCMC algorithms.

FPCA model				
n	VMP	MCMC		
10	0.66 (0.06)	0.62 (0.03)		
50	0.65 (0.07)	0.62 (0.04)		
100	0.66 (0.07)	0.62 (0.04)		
250	0.65 (0.06)	0.62 (0.04)		
500	0.65 (0.07)	0.62 (0.04)		
MlFPCA model				
n	VMP ($\boldsymbol{\zeta}^{(1)}$)	MCMC ($\boldsymbol{\zeta}^{(1)}$)	VMP ($\boldsymbol{\zeta}^{(2)}$)	MCMC ($\boldsymbol{\zeta}^{(2)}$)
10	0.45 (0.10)	0.40 (0.09)	0.64 (0.10)	0.56 (0.08)
50	0.43 (0.09)	0.39 (0.08)	0.62 (0.09)	0.56 (0.11)
100	0.45 (0.10)	0.41 (0.10)	0.63 (0.11)	0.58 (0.09)
250	0.42 (0.08)	0.40 (0.07)	0.62 (0.09)	0.56 (0.09)
500	0.44 (0.08)	0.40 (0.08)	0.63 (0.10)	0.57 (0.07)

Table 1: Median (median absolute deviation) RMSE for the scores.

6.2 Computational Speed Comparisons

In the previous section, we saw that the mean field product restriction in (3.1) does not compromise the accuracy of variational Bayesian inference for FPCA and similarly for MlFPCA. Another major advantage offered by variational Bayesian inference via VMP is fast approximate inference in comparison to MCMC simulations. Several published articles have confirmed the superior computational speed of variational Bayesian inference algorithms over MCMC simulations in numerous Bayesian models (Faes et al., 2011; Luts and Wand, 2015; Lee and Wand, 2016; Nolan et al., 2020).

A speed comparison of these algorithms is dependent on a number of factors related to the MCMC settings. In particular, Gibbs sampling can be used to address models (2.6) and (4.2), rather than programming through RStan. Nevertheless, our aim is to show that VMP for FPCA is significantly faster than readily available MCMC software that is commonly used in statistical applications (e.g. Goldsmith et al., 2015). While we compare to off-the-shelf MCMC software, variational Bayes is notably faster than tailored Gibbs samplers in related functional data settings (e.g. Goldsmith and Kitago, 2016). Other MCMC settings, such as the number of burn-in samples and MCMC iterations, were tuned throughout preliminary simulations to ensure computational accuracy.

In Table 2, we present a similar set of results for the computational speed of VMP and MCMC for models (2.6) and (4.2). The simulations were identical to those that were used to generate the results in Figure 3. In Table 2, we present the median elapsed computing time (in seconds), with the median absolute deviation in brackets. For the FPCA results, notice that most of the VMP simulations are completed within one minute, whereas the elapsed computing times for the MCMC simulations can exceed one hour for $n = 500$. The most impressive results are in the fourth column, where the median VMP simulations are shown to be roughly 25–100 times faster, depending on the values of n. A similar set of results are observed for the Bayesian MlFPCA model.

In preliminary analyses, we conducted an identical set of simulations, but set $L = 2$. Under these conditions, the VMP algorithm was 19.6 times faster than MCMC simulations for $n = 10$, 34.3 times faster for $n = 50$, 37.9 times faster for $n = 100$, 49.0 times faster for $n = 250$ and 59.8 times faster for $n = 500$. Comparing these results with those in column 4 of Table 2 of the FPCA panel, we see that the speed gains over MCMC simulations become more impressive as L increases.

The speed of the VMP algorithm is dependent on the convergence of the lower bound on the marginal log-likelihood (see Appendix E.1). These results are summarized in column 5 of Table 2, where the median number of iterations (with median absolute deviation in brackets) are presented. Most of the FPCA simulations converged within 150 iterations. Similarly, most of the MlFPCA simulations converged within 250 iterations.

6.3 Comparisons against a Covariance Decomposition Method

We now make a comparison of our Bayesian FPCA methodology against a conventional covariance decomposition method. Yao et al. (2005) present a nonparametric method for performing FPCA on sparse and irregular functional data. Their methodology, named principal components analysis through conditional expectation (PACE), is based on a covariance decomposition method and is available via the function FPCA() in the fdapace package (Chen et al., 2019) in R. We use an identical simulation setup to Section 6.1, but restrict n to 100.

The results for the accuracy in the estimation of the eigenfunctions is presented in Table 3. Estimation of the first three eigenfunctions is similar for both methods, while the estimation of the fourth eigenfunction is stronger through PACE. This effect was

FPCA model				
n	VMP	MCMC	MCMC/VMP	VMP Iterations
10	4.9 (2.0)	122.6 (17.7)	24.9	130 (64)
50	8.1 (3.1)	310.0 (30.9)	38.2	51 (22)
100	15.6 (4.7)	609.3 (34.8)	39.2	58 (18)
250	32.0 (7.4)	2564.7 (570.5)	80.0	46 (9)
500	59.3 (15.9)	5841.8 (889.6)	98.5	41 (9)
MlFPCA model				
n	VMP	MCMC	MCMC/VMP	VMP Iterations
10	28.3 (16.9)	1771.6 (247.2)	62.7	104 (79)
50	211.8 (83.1)	13789.0 (1194.0)	65.1	120 (91)
100	453.9 (205.8)	30938.9 (5250.9)	68.2	223 (111)
250	975.0 (519.1)	69845.4 (11572.5)	71.6	196 (116)
500	2906.7 (1311.5)	216260.0 (70246.9)	74.4	211 (100)

Table 2: Median (median absolute deviation) elapsed computing time for conducting Bayesian FPCA and Bayesian MlFPCA. The fourth column presents the ratio of the median elapsed time for MCMC to the median elapsed time for VMP. The fifth column shows the median (median absolute deviation) number of VMP iterations prior to convergence.

Method	ψ_1	ψ_2	ψ_3	ψ_4
VMP	-4.6 (0.7)	-3.3 (0.8)	-2.3 (0.8)	0.1 (0.9)
PACE	-4.4 (0.7)	-3.5 (07)	-2.0 (0.8)	-1.6 (0.8)

Table 3: Median (median absolute deviation) of the log ISE in estimating the eigenfunctions by conducting FPCA via VMP and PACE. We set $n = 100$ and $L = 4$.

also observed in Figure 3, where inference on the fourth eigenfunction was weaker for small and moderately sized datasets.

The median elapsed computing time for PACE was 50.8 seconds with a median absolute deviation of 0.6 seconds. Comparing this result with the corresponding elapsed computing time of the VMP algorithm in Table 2 for $n = 100$, we see that there is a clear advantage in speed for the VMP approach over the PACE algorithm. The median RMSE for PACE in estimating the scores is 0.77 with a median absolute deviation of 0.04. The corresponding result for the VMP approach in Table 1 for $n = 100$ shows that estimation of the scores is also stronger through VMP-based FPCA.

7 Application: United States Temperature Data

We now provide an illustration of our methodology with an application to temperature data collected from various United States weather stations, which is available from the rnoaa package (Chamberlain et al., 2021) in R. The rnoaa package is an interface to the National Oceanic and Atmospheric Administration's climate data. The function

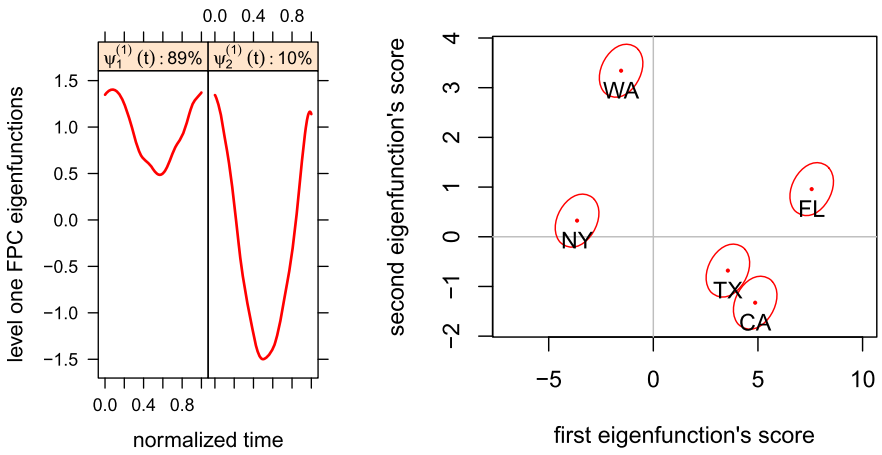

Figure 4: *Left panel*: The first two eigenfunctions of the first level covariance operator. The proportion of explained variability for each eigenfunction is also presented. *Right panel*: The corresponding scores for the weather stations in California (CA), Florida (FL), New York state (NY), Texas (TX) and Washington state (WA).

ghcnd_stations() provides access to all available global historical climatology network daily weather data for each weather site from 1960 to 1994. The information includes the longitude and latitude for each site, and this was used to determine the site's US state. Our analysis focused on maximum daily temperature over the 25 years of available data. From this package, we randomly selected a single weather station from each state and collected the weather station's daily maximum temperature recordings from 1960 to 1994. Additionally, we partition each weather station's dataset by year. The resulting dataset is a multilevel functional dataset, with the state (represented by a single weather station) as the first level and the year as the second level.

Chapter 8 of Ramsay and Silverman (2005) consider a similar example of Canadian temperature data from various weather stations. The difference in our analysis, aside from collecting US data, is the multilevel structure of the dataset. An additional consideration in the multilevel setting is the appropriate choices for L_1 and L_2, the number of retained eigenfunctions for the first and second level covariance operators. In the standard Bayesian FPCA model, we recommend initially setting $L = 15$ and truncating the number of retained eigenfunctions such that the proportion of explained variability is at least 95%. For the US temperature data, we apply this procedure to both levels of the model. We retained four eigenfunctions for the first level and eight eigenfunctions for the second level.

The results for the first level eigenfunctions and scores are presented in Figure 4. The first eigenfunction, which accounts for 89% of the first level variability, is a mean shift (since it is always positive). Its effect is stronger in the Winter months, indicating that US temperature is most variable in the Winter. Similar analysis of the second eigenfunction, which accounts for 10% of the total variability, shows that it represents

uniformity in the measured temperatures. It has positive contributions in the Winter months and negative contributions in the Summer months. As a consequence, weather stations at locations with larger discrepancies between Winter and Summer temperatures will have a strong and negative score for this eigenfunction. In the right panel of Figure 4, we present the scores for the weather stations in California (CA), Florida (FL), New York state (NY), Texas (TX) and Washington state (WA), as well as their 95% posterior credible boundaries. Florida, California and Texas all have positive scores for the first eigenfunction, indicating yearly maximal temperature recordings higher than the national average. New York and Washington have negative scores for the first eigenfunction, which is indicative of their lower than average temperatures. The scores for the second eigenfunction indicate that the greatest variability between Summer and Winter months can be found in California, whereas Washington state tends to have more uniform yearly temperature recordings. The second level eigenfunctions were mostly periodic and difficult to interpret so we have presented them in Appendix F.

8 Closing Remarks

We have provided a comprehensive overview of Bayesian FPCA and MlFPCA with a VMP-based mean field variational Bayes approach. Our coverage has focused on the Gaussian likelihood specification for the observed data, and it includes the introduction of three new fragments. In addition, a sequence of post-processing steps have been established to satisfy the orthogonality requirements of FPCA. There are numerous extensions that cannot be included in a single article including non-Gaussian likelihood specifications, multivariate FPCA modelling and experimentation with other spline or wavelet families for nonparametric regression. This article provides a clear means for resolving such methodological extensions.

Supplementary Material

Supplementary Material for Bayesian Functional Principal Components Analysis via Variational Message Passing with Multilevel Extensions (DOI: 10.1214/23-BA1393SUPP; .pdf).
Appendix A: Proof of Theorem 2.1. Detailed proofs of Theorem 2.1, Lemma 5.1 and Proposition 5.1.
Appendix B: Exponential Family Form. An overview of the exponential family forms for the normal and inverse-χ^2 density functions.
Appendix C: Algorithmic Derivations. Derivations of the VMP algorithmic updates in Algorithms 1, 2 and 3.
Appendix D: Multilevel Orthogonal Decomposition. An outline of the post-VMP steps for obtaining orthonormal eigenfunctions and uncorrelated scores at both levels of MlFPCA model.
Appendix E: Convergence and Algorithmic Updates. A description of the convergence of the VMP algorithms and pseudocode for the Bayesian FPCA and MlFPCA models.
Appendix F: Additional Experimental Results. A collection of additional experimental results.

References

Benko, M., Härdle, W., and Kneip, A. (2009). "Common functional principal components." *The Annals of Statistics*, 37: 1–34. MR2488343. doi: https://doi.org/10.1214/07-AOS516. 17

Bishop, C. M. (1999). "Variational Pincipal Components." In *Proceedings of the Ninth International Conference on Artificial Neural Networks*. Institute of Electrical and Electronics Engineers. 2

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer. 8

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational inference: A review for statisticians." *Journal of the American Statistical Association*, 112: 859–877. MR3671776. doi: https://doi.org/10.1080/01621459.2017.1285773. 3, 7

Chamberlain, S., Anderson, B., Salmon, M., Erickson, A., Potter, N., Stachelek, J., Simmons, A., Ram, K., Edmund, H., and rOpenSci (2021). "'NOAA' Weather Data from R." R package version 1.3.0. URL https://docs.ropensci.org/rnoaa/ 22

Chen, Y., Carroll, C., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., and Ji, H. (2019). "Principal Analysis by Conditional Expectation and Applications in Functional Data Analysis." R package version 0.5.9. URL https://cran.r-project.org/web/packages/fdapace/index.html 21

Di, C. Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). "Multilevel functional principal component analysis." *The Annals of Applied Statistics*, 3: 458–488. MR2668715. doi: https://doi.org/10.1214/08-AOAS206. 2, 5, 11, 12

Durbán, M., Harezlak, J., Wand, M. P., and Carroll, R. J. (2005). "Simple fitting of subject specific curves for longitudinal data." *Statistics in Medicine*, 24: 1153–1167. MR2134571. doi: https://doi.org/10.1002/sim.1991. 6

Faes, C., Ormerod, J. T., and Wand, M. P. (2011). "Variational Bayesian inference for parametric and nonparametric regression with missing data." *Journal of the American Statistical Association*, 106: 959–971. MR2894756. doi: https://doi.org/10.1198/jasa.2011.tm10301. 20

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1: 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 7

Gentle, J. E. (2007). *Matrix Algebra*. New York: Springer. MR2337395. doi: https://doi.org/10.1007/978-0-387-70873-7. 4

Goldsmith, J. and Kitago, T. (2016). "Assessing systematic effects of stroke on motor-control by using hierarchical function-on-scalar regression." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65: 215–236. MR3456686. doi: https://doi.org/10.1111/rssc.12115. 21

Goldsmith, J. and Schwartz, J. E. (2017). "Variable selection in the functional linear

concurrent model." *Statistics in Medicine*, 36: 2237–2250. MR3660128. doi: https://doi.org/10.1002/sim.725. 3

Goldsmith, J., Zippunnikov, V., and Schrack, J. (2015). "Generalized multilevel function-on-scalar regression and principal component analysis." *Biometrics*, 71: 344–353. MR3366239. doi: https://doi.org/10.1111/biom.12278. 3, 7, 12, 21

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). "Principal component models for sparse functional data." *Biometrika*, 87: 587–602. MR1789811. doi: https://doi.org/10.1093/biomet/87.3.587. 2, 3

Lee, C. Y. Y. and Wand, M. P. (2016). "Streamlined mean field variational Bayes for longitudinal and multilevel data analysis." *Biometrical Journal*, 58: 868–895. MR3527420. doi: https://doi.org/10.1002/bimj.201500007. 20

Luts, J. and Wand, M. P. (2015). "Variational inference for count response semiparametric regression." *Bayesian Analysis*, 10: 991–1023. MR3432247. doi: https://doi.org/10.1214/14-BA932. 20

Maestrini, L. and Wand, M. P. (2021). "The Inverse G-Wishart distribution and variational message passing." *Australian and New Zealand Journal of Statistics*, 63: 517–541. MR4374511. doi: https://doi.org/10.1111/anzs.12339. 7, 9

Menictas, M. and Wand, M. P. (2013). "Variational inference for marginal longitudinal semiparametric regression." *Stat*, 2: 61–71. MR4027301. doi: https://doi.org/10.1002/sta4.18. 8

Minka, T. (2005). "Divergence measures and message passing." Technical report, Microsoft Research Ltd., Cambridge, UK. 3, 7, 8

Nolan, T. H. (2020). "Variational Bayesian inference: message passing schemes and streamlined multilevel data analysis." Ph.D. thesis, University of Technology Sydney. 3

Nolan, T. H., Goldsmith, J., and Ruppert, D. (2023). "Supplementary Material for "Bayesian Functional Principal Components Analysis via Variational Message Passing with Multilevel Extensions"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1393SUPP. 5

Nolan, T. H., Menictas, M., and Wand, M. P. (2020). "Streamlined computing for variational inference with higher level random effects." *Journal of Machine Learning Research*, 21: 1–62. MR4209443. 15, 20

Nolan, T. H. and Wand, M. P. (2020). "Streamlined Solutions to Multilevel Sparse Matrix Problems." *ANZIAM Journal*, 62: 18–41. MR4130820. doi: https://doi.org/10.1017/s1446181120000061. 15

Ormerod, J. T. and Wand, M. P. (2010). "Explaining variational approximations." *The American Statistician*, 64: 140–153. MR2757005. doi: https://doi.org/10.1198/tast.2010.09058. 3, 7, 8, 9

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer. MR2168993. 2, 23

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ 18

Ruppert, D. (2002). "Selecting the Number of Knots for Penalized Splines." *Journal of Computational and Graphical Statistics*, 1: 735–757. MR1944261. doi: https://doi.org/10.1198/106186002321018768. 19

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press. MR1998720. doi: https://doi.org/10.1017/CBO9780511755453. 6

Ruppert, D., Wand, M. P., and Carroll, R. J. (2009). "Semiparametric regression during 2003–2007." *Electronic Journal of Statistics*, 3: 1193–1256. MR2566186. doi: https://doi.org/10.1214/09-EJS525. 6

Stan Development Team (2020). "RStan: the R interface to Stan." R package version 2.21.2. URL http://mc-stan.org/ 3, 18

Tipping, M. E. and Bishop, C. M. (1999). "Probabilistic principal component analysis." *Journal of the Royal Statistical Society, Series B*, 3: 611–622. MR1707864. doi: https://doi.org/10.1111/1467-9868.00196. 2, 7

van der Linde, A. (2008). "Variational Bayesian functional PCA." *Computational Statistics and Data Analysis*, 53: 517–533. MR2649106. doi: https://doi.org/10.1016/j.csda.2008.09.015. 2, 3, 7

Wand, M. P. (2017). "Fast approximate inference for arbitrarily large semiparametric regression models via message passing (with discussion)." *Journal of the American Statistical Association*, 112: 137–168. MR3646558. doi: https://doi.org/10.1080/01621459.2016.1197833. 3, 7, 8, 9, 13

Wand, M. P. and Ormerod, J. T. (2008). "On semiparametric regression with O'Sullivan penalized splines." *Australian & New Zealand Journal of Statistics*, 50: 179–198. MR2431193. doi: https://doi.org/10.1111/j.1467-842X.2008.00507.x. 6

Wang, J. L., Chiou, J. M., and Müller, H. G. (2016). "Functional data analysis." *Annual Review of Statistics and Its Applications*, 3: 257–295. 2

Winn, J. and Bishop, C. M. (2005). "Variational message passing." *Journal of Machine Learning Research*, 6: 661–694. MR2249835. 3

Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2016). "Fast covariance estimation for high-dimensional functional data." *Statistics and computing*, 26(1-2): 409–421. MR3439382. doi: https://doi.org/10.1007/s11222-014-9485-x. 5

Yao, F., Müller, H. G., and Wang, J. L. (2005). "Functional data analysis for sparse longitudinal data." *Journal of the American Statistical Association*, 100: 577–590. MR2160561. doi: https://doi.org/10.1198/016214504000001745. 2, 4, 5, 17, 21