Bayesian Functional Principal Components Analysis via Variational Message Passing

Tui H. Nolan *[1,2,3], Jeff Goldsmith[4], and David Ruppert[1,5]

[1]School of Operations Research and Information Engineering, Cornell University

[2]Medical Research Council Biostatistics Unit, The University of Cambridge

[3]School of Mathematical and Physical Sciences, University of Technology Sydney

[4]Department of Biostatistics, Mailman School of Public Health, Columbia University

[5]Department of Statistics and Data Science, Cornell University

July 9, 2021

*Corresponding author: tn352@cam.ac.uk

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. Indeed, this is the first article to address a functional data model via variational message passing. Our approach introduces two 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

1 Introduction

Functional principal components analysis (FPCA) is the methodological extension of classical principal components analysis (PCA) to functional data. Within the overarching framework of functional data analysis, FPCA is a central technique. 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 realisations 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 the reduced set of eigenfunctions.

There are technical issues that arise in the functional setting that are not present for multivariate data. The domain of the functional curves is typically a compact interval $[0,T]$ of the real line. Despite having a continuous domain, the curves are only observed at discrete points over this interval. Furthermore, the points of observation, as well as the total number of observations, need not be the same for each curve. 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 from the perspective of integration, rather than inner products of vectors.

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, Chiou, and Müller (2016, Section 2). Much of this work mirrors the eigendecomposition approach to PCA, in that an eigenbasis is obtained from a covariance surface. Yao, Müller, and Wang (2005) focused on the case of sparsely observed functional data, and estimate principal component scores through conditional expectations. Xiao, Zipunnikov, Ruppert, and Crainiceanu

(2016) developed a fast covariance estimation method for densely observed functional data. Di, Crainiceanu, Caffo, and Punjabi (2009) extended FPCA to multilevel functional data, extracting within and between subject sources of variability, and Greven, Crainiceanu, Caffo, and Reich (2011) developed methods for longitudinal functional data. However, Goldsmith, Greven, and Crainiceanu (2013) noted that these approaches implicitly condition on an estimated eigenbasis to estimate scores, meaning that inference on individual curve estimates can be inaccurate.

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 covariance and then estimating scores, all quantities are considered unknown and are estimated jointly. James, Hastie, and Sugar (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. Goldsmith, Zippunnikov, and Schrack (2015) introduced a fully Bayes framework for multilevel function-on-scalar regression models with FPCA applied to two levels of residuals, and also considered observed values that arise from exponential family distributions.

In frequentist 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. The key advantage in the Bayesian approach is that the covariance function is not estimated, meaning that complex bivariate smoothing is not required. Indeed, 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 for dense functional data, and for sparse, irregular functional data – where estimating the raw covariance is difficult or impossible – direct estimation of eigen-

functions in a Bayesian model 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 is a computationally expensive procedure and, in some biostatistical applications (Goldsmith et al., 2015), the computational time can reach several hours. Second, current versions of variational Bayes for FPCA, despite being a much faster computational alternative, are difficult to extend to more complex likelihood specifications, such as multilevel data models and binary response outcomes.

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 & Wand, 2010; Blei, Kucukelbir, & McAuliffe, 2017) can be incorporated into this framework through an alternate scheme known as variational message passing (VMP) (Winn & 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 propose an FPCA extension of the VMP framework for variational Bayesian inference set out in Wand (2017). Our novel methodology includes the introduction of two fragments that are necessary for computing approximate posterior density functions under an MFVB scheme, 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. In Section 4, we outline the post-VMP steps that are required for producing orthonormal eigenfunctions. Simula-

tions, including speed and accuracy comparisons with MCMC algorithms, are presented in Section 5, and an application to United States temperature data is provided in Section 6.

2 Functional Principal Components Analysis

Consider a random sample of i.i.d. smooth random functions $y_1, \ldots, y_n \in L^2[0,1]$. We will assume the existence of a continuous mean function $\mu = \mathbb{E} y_i$ and continuous covariance surface $\sigma(t,s) = \mathbb{E}[\{y_i(t) - \mu(t)\}\{y_i(s) - \mu(s)\}]$, $i = 1, \ldots, n$. Then, the covariance operator Σ of y_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):

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

where $\zeta_{il}^* = \int_0^1 \{y_i(t) - \mu(t)\} \psi_l^*(t) dt$ are the principal components scores. The ζ_{il}^* are independent across i and uncorrelated across l, with $\mathbb{E}(\zeta_{il}^*) = 0$ and $\mathbb{V}\text{ar}(\zeta_{il}^*) = \gamma_l$. Note that under the Gaussian assumptions that we introduce in (6), the ζ_{il}^* are also independent across l. The asterisk is used as a reminder that the eigenfunctions in (1) are orthonormal and that the scores are independent.

Expansion (1) facilitates dimension reduction by providing a best approximation for each curve $y_1, \ldots, y_n$ 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 $\psi_1, \ldots, \psi_L$, the minimum of $\sum_{i=1}^n \left|\left| y_i - \mu - \sum_{l=1}^L \langle y_i - \mu, \psi_l \rangle \psi_l \right|\right|^2$ is achieved for $\psi_l = \psi_l^*$, $l = 1, \ldots, L$, 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 y_i as

$$\widehat{y}_i(t) \equiv \mu(t) + \sum_{l=1}^{L} \zeta_{il}^* \, \psi_l^*(t), \quad i = 1, \ldots, n. \tag{2}$$

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) are also possible, where

$$\widehat{y}_i(t) \equiv \mu(t) + \sum_{l=1}^{L} \zeta_{il} \, \psi_l(t), \quad i = 1, \ldots, n, \tag{3}$$

where ζ_{il} are correlated across l, but remain independent across i, and the ψ_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) from (3). Its proof is provided in Appendix A.

Theorem 2.1. *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* $\widehat{y}_i(t) = \mu(t) + \sum_{l=1}^{L} \zeta_{il}^* \, \psi_l^*(t)$.

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 $y_1, \ldots, y_n$ are indirectly observed as noisy observations at discrete points in time. Furthermore, the observation times are not necessarily identical for each curve. Let the set of design points for the ith curve be summarized by the vector $\boldsymbol{t}_i \equiv (t_{i1}, \ldots, t_{iT_i})^\intercal$ and the observations for the ith curve, $y_i(t)$, by the vector $\boldsymbol{y}_i \equiv \{y_i(t_{i1}) + \varepsilon_{i1}, \ldots, y_i(t_{iT_i}) + \varepsilon_{iT_i}\}^\intercal$, where T_i is the number of observations on the ith curve and ε_{ij} are i.i.d. noise terms with $\mathbb{E}(\varepsilon_{ij}) = 0$ and $\mathbb{V}\text{ar}(\varepsilon_{ij}) = \sigma^2_\varepsilon$. The finite decomposition in (2) takes the form:

$$\boldsymbol{y}_i = \boldsymbol{\mu}_i + \sum_{l=1}^{L} \zeta_{il}\boldsymbol{\psi}_{il} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{4}$$

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

We model continuous curves from discrete observations via nonparametric regression (Ruppert, Wand, & Carroll, 2003, 2009), using the mixed model-based penalized spline basis function representation, as in Durbán, Harezlak, Wand, and Carroll (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 $\{z_k(\cdot)\}_{1 \le k \le K}$ 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 described in Section 4 of Wand and Ormerod (2008).

In order to avoid notational clutter, we incorporate the following definitions: $\boldsymbol{\beta}_\mu \equiv$

$(\beta_{\mu,0}, \beta_{\mu,1})^\intercal$, $\boldsymbol{u}_\mu \equiv (u_{\mu,1}, \ldots, u_{\mu,K})^\intercal$, $\boldsymbol{\nu}_\mu \equiv (\boldsymbol{\beta}_\mu^\intercal, \boldsymbol{u}_\mu^\intercal)^\intercal$, where $\boldsymbol{\nu}_\mu$ is the vector of spline coefficients for $\mu(t)$; and $\boldsymbol{\beta}_{\psi_l} \equiv (\beta_{\psi_l,0}, \beta_{\psi_1,1})^\intercal$, $\boldsymbol{u}_{\psi_l} \equiv (u_{\psi_l,1}, \ldots, u_{\psi_l,K})^\intercal$ and $\boldsymbol{\nu}_{\psi_l} \equiv (\boldsymbol{\beta}_{\psi_l}^\intercal, \boldsymbol{u}_{\psi_l}^\intercal)^\intercal$ for $l = 1, \ldots, L$, where $\boldsymbol{\nu}_{\psi_l}$ is the vector of spline coefficients for $\psi_l(t)$. Then simple derivations that stem from (4) show that the vector of observations on each of the response curves satisfies the representation $\boldsymbol{y}_i = \boldsymbol{C}_i(\boldsymbol{\nu}_\mu + \sum_{l=1}^{L} \zeta_{il}\boldsymbol{\nu}_{\psi_l}) + \boldsymbol{\epsilon}_i$, where

$$\boldsymbol{C}_i \equiv \begin{bmatrix} 1 & t_{i1} & z_1(t_{i1}) & \ldots & z_K(t_{i1}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{iT_i} & z_1(t_{iT_i}) & \ldots & z_K(t_{iT_i}) \end{bmatrix}. \tag{5}$$

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

Next, we present the Bayesian FPCA Gaussian response model:

$$\begin{gathered}
\boldsymbol{y}_i | \boldsymbol{\nu}, \boldsymbol{\zeta}_i, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} \mathrm{N}\left\{ \boldsymbol{C}_i \left(\boldsymbol{\nu}_\mu + \sum_{l=1}^{L} \zeta_{il}\boldsymbol{\nu}_{\psi_l} \right), \sigma_\varepsilon^2 \boldsymbol{I}_{T_i} \right\}, \quad \boldsymbol{\zeta}_i \overset{\text{ind.}}{\sim} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\zeta_i}), \quad i = 1, \ldots, n, \\
\begin{bmatrix} \boldsymbol{\nu}_\mu \\ \boldsymbol{\nu}_{\psi_l} \end{bmatrix} \Bigg| \; \sigma_\mu^2, \sigma_{\psi_l}^2 \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_\mu \\ \boldsymbol{\mu}_{\psi_l} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_\mu & \boldsymbol{O}^\intercal \\ \boldsymbol{O} & \boldsymbol{\Sigma}_{\psi_l} \end{bmatrix} \right), \quad \sigma_{\psi_l}^2 | a_{\psi_l} \overset{\text{ind.}}{\sim} \text{Inverse} - \chi^2(1, 1/a_{\psi_l}), \\
a_{\psi_l} \overset{\text{ind.}}{\sim} \text{Inverse} - \chi^2(1, 1/A_{\psi_l}^2), \quad l = 1, \ldots, L, \\
\sigma_\mu^2 | a_\mu \sim \text{Inverse} - \chi^2(1, 1/a_\mu), \quad a_\mu \sim \text{Inverse} - \chi^2(1, 1/A_\mu^2), \\
\sigma_\varepsilon^2 | a_\varepsilon \sim \text{Inverse} - \chi^2(1, 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse} - \chi^2(1, 1/A_\varepsilon^2),
\end{gathered} \tag{6}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{\mu} &\equiv (\boldsymbol{\mu}_{\beta_\mu}^{\intercal}, \mathbf{0}_K^{\intercal})^{\intercal}, \quad \boldsymbol{\Sigma}_{\mu} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\beta_\mu} & \mathbf{O}^{\intercal} \\ \mathbf{O} & \sigma_\mu^2 \boldsymbol{I}_K \end{bmatrix}, \\
\boldsymbol{\mu}_{\psi_l} &\equiv (\boldsymbol{\mu}_{\beta_{\psi_l}}^{\intercal}, \mathbf{0}_K^{\intercal})^{\intercal}, \quad \boldsymbol{\Sigma}_{\psi_l} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\beta_{\psi_l}} & \mathbf{O}^{\intercal} \\ \mathbf{O} & \sigma_{\psi_l}^2 \boldsymbol{I}_K \end{bmatrix}, \quad l = 1, \ldots, L,
\end{aligned} \tag{7}$$

and $\boldsymbol{\mu}_{\beta_\mu}$ (2×1), $\boldsymbol{\mu}_{\beta_{\psi_l}}$ $(2 \times 1,\ l = 1, \ldots, L)$, $\boldsymbol{\Sigma}_{\beta_\mu}$ $(2 \times 2,$ positive definite$)$, $\boldsymbol{\Sigma}_{\beta_{\psi_l}}$ $(2 \times 2,$ positive definite, $l = 1, \ldots, L)$, $\boldsymbol{\Sigma}_{\zeta_i}$ $(L \times L,$ positive definite, $i = 1, \ldots, n)$, $A_\nu > 0$, $A_{\psi_l} > 0$ $(l = 1, \ldots, L)$ are the model hyperparameters. Note that the iterated inverse-χ^2 distributional specification on σ_ε^2, which involves an inverse-χ^2 prior specification on the auxiliary variable a_ε, is equivalent to $\sigma_\varepsilon^2 \sim \text{Half-Cauchy}(A_\varepsilon)$. 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_μ^2 and $\sigma_{\psi_1}^2, \ldots, \sigma_{\psi_L}^2$.

3 Variational Bayesian Inference

In keeping with the theme of this article, we will explain variational Bayesian inference and its extensions to variational message passing in the context of the Bayesian FPCA model (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 variational message passing.

Full Bayesian inference for the parameter set $\boldsymbol{\nu}$, $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$, σ_ε^2, a_ε, σ_μ^2, a_μ, $\sigma_{\psi_1}^2, \ldots, \sigma_{\psi_L}^2$ and $a_{\psi_1}, \ldots, a_{\psi_L}$ requires the determination of the posterior density function $p(\boldsymbol{\nu}, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n, \sigma_\varepsilon^2, a_\varepsilon, \sigma_\mu^2, a_\mu, \sigma_{\psi_1}^2, \ldots, \sigma_{\psi_L}^2, a_{\psi_1}, \ldots, a_{\psi_L} | \boldsymbol{y})$, but it is typically analytically intractable. The standard approach for overcoming this deficiency is to employ MCMC approaches. However, we propose two major arguments against this approach. First, MCMC simulations

are very slow for model (6), even for moderate dimensions of $\boldsymbol{\nu}$, which depends on the number of eigenfunctions (L) and O'Sullivan penalized spline basis functions (K). Second, the mean function $\mu(t)$ and the eigenfunctions $\psi_1(t), \ldots, \psi_L(t)$ are typically highly correlated, which is expected to lead to poor mixing. A possible remedy for this is to use an inverse G-Wishart prior structure that permits correlations amongst the spline coefficients (Goldsmith & Kitago, 2016). However, this is beyond the scope of this article, which is not concerned with improving MCMC methods for FPCA.

Alternatively, variational approximate inference for model (6) involves the mean field restriction:

$$p(\boldsymbol{\nu}, \zeta_1, \ldots, \zeta_n, \sigma^2_\varepsilon, a_\varepsilon, \sigma^2_\mu, a_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}, a_{\psi_1}, \ldots, a_{\psi_L} | \boldsymbol{y}) \approx \left\{ \prod_{i=1}^{N} q(\zeta_i) \right\} q(\boldsymbol{\nu}, a_\varepsilon, a_\mu, a_{\psi_1}, \ldots, a_{\psi_L}) q(\sigma^2_\varepsilon, \sigma^2_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}). \tag{8}$$

where each q represents an approximate density function. The q-density functions are selected to minimize the Kullback-Liebler divergence of the left-hand side of (8) from its right-hand side. The approximation in (8) represents the minimal mean-field restriction that is required for approximate variational inference. Here, we have assumed posterior independence between global parameters (spline coefficients for the mean curve and the eigenfunctions) and response curve-specific parameters (the scores), as well as incorporating the notion of *asymptotic independence* between regression coefficients and variance parameters (Menictas & Wand, 2013, Section 3.1). However, induced factorizations, based on graph theoretic results (Bishop, 2006, Section 10.2.5), admit further factorizations, and the right-hand side of (8) becomes

$$\left\{\prod_{i=1}^{N} q(\zeta_i)\right\} q(\boldsymbol{\nu}) q(\sigma_\varepsilon^2) q(a_\varepsilon) q(\sigma_\mu^2) q(a_\mu) \left\{\prod_{l=1}^{L} q(\sigma_{\psi_l}^2) q(a_{\psi_l})\right\}. \tag{9}$$

From here, we work with the factorization in (9) to minimize the Kullback-Leibler divergence of the right-hand side of (8) from its left-hand side.

The parameter vectors that define each of the q-density functions are interrelated and are updated by a coordinate ascent algorithm (Ormerod & 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. For instance, the updates for the optimal posterior density functions of the coefficients in a linear regression model will differ from those in a linear logistic regression 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 (6) that represents the factorization in (9) is presented in Figure 1. Each probability density specification in (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 (8) from its right-hand side are referred to as optimal q-density functions.

Our presentation of the variational message passing construction will focus on computing the optimal q-density functions for $\boldsymbol{\nu}$ and $\zeta_1, \ldots, \zeta_n$. As explained in Minka (2005), the

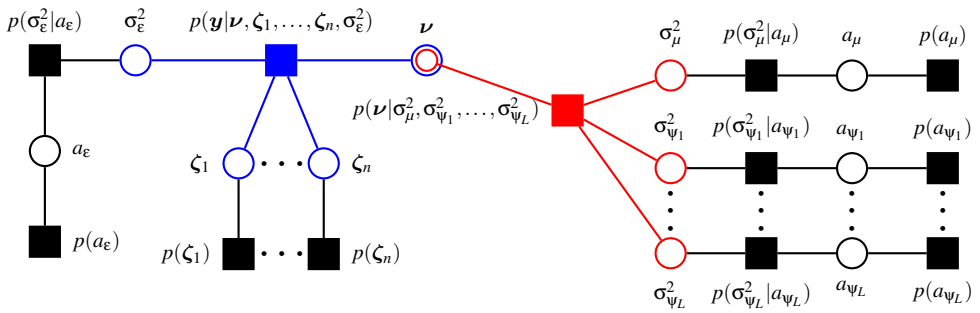


Figure 1: The factor graph for the Bayesian FPCA model in (6).

q-density function for $\boldsymbol{\nu}$ and $\zeta_1,\ldots,\zeta_n$ can be expressed as

$$\begin{aligned} q(\boldsymbol{\nu}) &\propto m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\to\boldsymbol{\nu}}(\boldsymbol{\nu})\; m_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\to\boldsymbol{\nu}}(\boldsymbol{\nu}) \\ q(\zeta_i) &\propto m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\to\zeta_i}(\zeta_i)\; m_{p(\zeta_i)\to\zeta_i}(\zeta_i), \quad i=1,\ldots,n. \end{aligned} \tag{10}$$

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).

3.1.1 Exponential Family Form

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 (10) are typically in the exponential family of density functions. Wand (2017) explains how natural parameter vectors play a 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 (10) take the form

$$\begin{aligned}\boldsymbol{\eta}_{q^*(\boldsymbol{\nu})} &= \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\boldsymbol{\nu}} \\ \boldsymbol{\eta}_{q^*(\zeta_i)} &= \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\zeta_i} + \boldsymbol{\eta}_{p(\zeta_i)\to\zeta_i}, \quad i = 1,\ldots,n.\end{aligned} \tag{11}$$

Before introducing the exponential family forms for key distributions in the VMP setting, we outline some matrix and vector operators. 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 $\boldsymbol{D}_d^+\,\text{vec}(\boldsymbol{A}) = \text{vech}(\boldsymbol{A})$.

Now we describe the exponential family form for the normal distribution. For a $d \times 1$ multivariate normal random vector $\boldsymbol{x} \sim \text{N}(\boldsymbol{\mu}, \Sigma)$, the probability density function of $\boldsymbol{x}$ can be shown to satisfy

$$p(\boldsymbol{x}) = \exp\left\{\boldsymbol{T}_{\text{vec}}(\boldsymbol{x})^\intercal \boldsymbol{\eta}_{\text{vec}} - A_{\text{vec}}(\boldsymbol{\eta}_{\text{vec}}) - \frac{d}{2}\log(2\pi)\right\}, \tag{12}$$

where $\boldsymbol{T}_{\text{vec}}(\boldsymbol{x}) \equiv \{\boldsymbol{x}^\intercal, \text{vec}(\boldsymbol{x}\boldsymbol{x}^\intercal)^\intercal\}^\intercal$ is the vector of sufficient statistics and $\boldsymbol{\eta}_{\text{vec}} \equiv (\boldsymbol{\eta}_{\text{vec},1}^\intercal, \boldsymbol{\eta}_{\text{vec},2}^\intercal)^\intercal \equiv [(\Sigma^{-1}\boldsymbol{\mu})^\intercal, -\frac{1}{2}\{\text{vec}(\Sigma^{-1})\}^\intercal]^\intercal$ is the natural parameter vector. The function $A_{\text{vec}}(\boldsymbol{\eta}_{\text{vec}}) = -\frac{1}{4}\boldsymbol{\eta}_{\text{vec},1}^\intercal\{\text{vec}^{-1}(\boldsymbol{\eta}_{\text{vec},2})\}^{-1}\boldsymbol{\eta}_{\text{vec},1} - \frac{1}{2}\log|-2\,\text{vec}^{-1}(\boldsymbol{\eta}_{\text{vec},2})|$ is the log-partition function. The inverse mapping of the natural parameter vector is (Wand, 2017, equation S.4)

$$\boldsymbol{\mu} = -\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{\eta}_{\text{vec},2})\right\}^{-1}\boldsymbol{\eta}_{\text{vec},1} \quad \text{and} \quad \boldsymbol{\Sigma} = -\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{\eta}_{\text{vec},2})\right\}^{-1}. \tag{13}$$

We will refer to the representation of the multivariate normal probability density function in (12) as the *vec-based representation*.

Alternatively, a more storage-economical representation of the multivariate normal probability density function is the *vech-based representation*:

$$p(\boldsymbol{x}) = \exp\left\{\boldsymbol{T}_{\text{vech}}(\boldsymbol{x})^{\intercal}\boldsymbol{\eta}_{\text{vech}} - A_{\text{vech}}(\boldsymbol{\eta}_{\text{vech}}) - \frac{d}{2}\log(2\pi)\right\},$$

where the vector of sufficient statistics, the natural parameter vector and the log-partition function are, $\boldsymbol{T}_{\text{vech}}(\boldsymbol{x}) \equiv \{\boldsymbol{x}^{\intercal}, \text{vec}(\boldsymbol{x}\boldsymbol{x}^{\intercal})^{\intercal}\}^{\intercal}$, $\boldsymbol{\eta}_{\text{vech}} \equiv (\boldsymbol{\eta}_{\text{vech},1}^{\intercal}, \boldsymbol{\eta}_{\text{vech},2}^{\intercal})^{\intercal} \equiv [(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{\intercal}, -\frac{1}{2}\boldsymbol{D}_d^{\intercal}\{\text{vec}(\boldsymbol{\Sigma}^{-1})\}^{\intercal}]^{\intercal}$ and $A_{\text{vech}}(\boldsymbol{\eta}_{\text{vech}}) = -\frac{1}{4}\boldsymbol{\eta}_{\text{vech},1}^{\intercal}\{\text{vec}^{-1}(\boldsymbol{D}_d^{+\intercal}\boldsymbol{\eta}_{\text{vech},2})\}^{-1}\boldsymbol{\eta}_{\text{vech},1} - \frac{1}{2}\log|-2\,\text{vec}^{-1}(\boldsymbol{D}_d^{+\intercal}\boldsymbol{\eta}_{\text{vech},2})|$, respectively. The inverse mapping of the natural parameter vector under the vech-based representation is

$$\boldsymbol{\mu} = -\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{D}_d^{+\intercal}\boldsymbol{\eta}_{\text{vech},2})\right\}^{-1}\boldsymbol{\eta}_{\text{vech},1} \quad \text{and} \quad \boldsymbol{\Sigma} = -\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{D}_d^{+\intercal}\boldsymbol{\eta}_{\text{vech},2})\right\}^{-1}. \tag{14}$$

The other major distribution within the exponential family that is pivotal for this article is the inverse-χ^2 distribution. A random variable x has an inverse-χ^2 distribution with shape parameter $\xi > 0$ and scale parameter $\lambda > 0$ if the probability density function of x is

$$p(x) = \frac{(\lambda/2)^{\xi/2}}{\Gamma(\xi/2)}x^{-(\xi+2)/2}\exp\left(-\frac{\lambda}{2x}\right)\mathbb{I}(x > 0),$$

where the vector of sufficient statistics, the natural parameter vector and the log-partition function are $\boldsymbol{T}(x) \equiv (\log(x), 1/x)^{\intercal}$, $\boldsymbol{\eta} = (\eta_1, \eta_2)^{\intercal} = \{-\frac{1}{2}(\xi+2), -\frac{\lambda}{2}\}^{\intercal}$ and $A(\boldsymbol{\eta}) \equiv \log\{\Gamma(\xi/2)\} -$

$\frac{\xi}{2}\log(\lambda/2)$, respectively. Note that $\Gamma(z) \equiv \int_0^\infty u^{z-1}e^u du$ is the gamma function, $\mathbb{I}(\cdot)$ is the indicator function, $\zeta > 0$ is the scale parameter and $\lambda > 0$ is the shape parameter. The inverse mapping of the natural parameter vector is $\xi = -2\eta_1 - 2$ and $\lambda = -2\eta_2$.

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 *functional principal component Gaussian penalization fragment* (red in Figure 1). The fragments for $p(\zeta_1), \ldots, p(\zeta_n)$ are *Gaussian prior fragments* (Wand, 2017, Section 4.1.1); the fragments for $p(\sigma_\varepsilon^2|a_\varepsilon)$, $p(\sigma_\mu^2|a_\mu)$ 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 & Wand, 2020, Algorithm 2); and $p(a_\varepsilon)$, $p(a_\mu)$ and $p(a_{\psi_1}), \ldots, p(a_{\psi_L})$ are univariate versions of the *inverse G-Wishart prior fragment* (Maestrini & Wand, 2020, Algorithm 1).

3.2 Functional Principal Component Gaussian Likelihood Fragment

The message from $p(\boldsymbol{y}|\boldsymbol{\nu}, \zeta_1, \ldots, \zeta_n, \sigma_\varepsilon^2)$ to $\boldsymbol{\nu}$ can be shown to be proportional to a multivariate normal density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\to\boldsymbol{\nu}} \longleftarrow \begin{bmatrix} \mathbb{E}_q(1/\sigma_\varepsilon^2)\displaystyle\sum_{i=1}^n \left\{\mathbb{E}_q(\widetilde{\zeta}_i)^\mathsf{T} \otimes \boldsymbol{C}_i\right\}^\mathsf{T} \boldsymbol{y}_i \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma_\varepsilon^2)\displaystyle\sum_{i=1}^n \operatorname{vec}\left\{\mathbb{E}_q(\widetilde{\zeta}_i\widetilde{\zeta}_i^\mathsf{T}) \otimes (\boldsymbol{C}_i^\mathsf{T}\boldsymbol{C}_i)\right\} \end{bmatrix}, \qquad (15)$$

where $\widetilde{\zeta}_i \equiv (1, \zeta_i^\mathsf{T})^\mathsf{T}, \quad i = 1, \ldots, n$. Note that the natural parameter vector in (15) is within the vec-based representation of a normal probability density function. In preliminary simulations, we found that computations using the vech-based representation were enormously hindered by the need to use a huge Moore-Penrose inverse matrix. For instance, consider the case where there are two basis functions ($L = 2$) and 25 O'Sullivan penalized spline basis functions ($K = 25$) for nonparametric regression. In this instance, the vector $\boldsymbol{\nu}$ is 81×1 ($d = 81$) and the Moore-Penrose inverse matrix $\boldsymbol{D}_{81}^+$ has dimension 3321×6561,

inhibiting the computational speed. For this reason, we have decided to use the vec-based representation for $m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\to\boldsymbol{\nu}}$, which does not require the use of a Moore-Penrose inverse matrix.

For each $i = 1,\ldots,n$, the message from $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)$ to ζ_i is proportional to a multivariate normal density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\to\zeta_i} \longleftarrow \begin{bmatrix} \mathbb{E}_q(1/\sigma_\varepsilon^2)\left\{\mathbb{E}_q(\boldsymbol{V}_\psi)^\intercal \boldsymbol{C}_i^\intercal \boldsymbol{y}_i - \mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i})\right\} \\ -\frac{1}{2}\mathbb{E}_q(1/\sigma_\varepsilon^2)\boldsymbol{D}_L^\intercal \operatorname{vec}\{\mathbb{E}_q(\boldsymbol{H}_{\psi,i})\} \end{bmatrix}, \tag{16}$$

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 $\boldsymbol{H}_{\psi,i} \equiv \boldsymbol{V}_\psi^\intercal \boldsymbol{C}_i^\intercal \boldsymbol{C}_i \boldsymbol{V}_\psi$. Note that this natural parameter vector is within the vech-based representation of a normal probability density function.

The message from $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)$ to σ_ε^2 is an inverse-χ^2 density function, with natural parameter vector

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\to\sigma_\varepsilon^2} \longleftarrow \begin{bmatrix} -\frac{1}{2}\sum_{i=1}^n T_i \\ -\frac{1}{2}\sum_{i=1}^n \mathbb{E}_q\left\{\left(\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{V}\widetilde{\boldsymbol{\zeta}}_i\right)^\intercal \left(\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{V}\widetilde{\boldsymbol{\zeta}}_i\right)\right\} \end{bmatrix}, \tag{17}$$

where $\boldsymbol{V} \equiv [\ \boldsymbol{\nu}_\mu\ \ \boldsymbol{\nu}_{\psi_1}\ \ \ldots\ \ \boldsymbol{\nu}_{\psi_L}\ ]$. Note that the inverse-$\chi^2$ density function message that is passed to σ_ε^2 is part of the inverse G-Wishart class of density functions. In keeping with the formalisms set out in Maestrini and Wand (2020), a graph message is also required for conjugate variational inference. According to Section 7.4 of Maestrini and Wand (2020), the auxiliary-based hierarchical prior specification of σ_ε^2 in (6) requires a graphical message of the form

$$G_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon} \longleftarrow G_{\text{full}}. \tag{18}$$

See Maestrini and Wand (2020) for further details on graphical parameters in inverse G-Wishart and inverse-χ^2 density functions.

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 D.

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

Inputs: $\boldsymbol{\eta}_{\boldsymbol{\nu}\to p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)}$, $\{\boldsymbol{\eta}_{\zeta_i\to p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)} : i = 1,\ldots,n\}$
$\{\boldsymbol{\eta}_{\sigma^2_\varepsilon\to p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)}, G_{\sigma^2_\varepsilon\to p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)}\}$

Updates:

Update posterior expectations. ▷ see Appendix D
Update $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}}$ ▷ equation (15)
for $i = 1,\ldots,n$ **do**
    Update $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\zeta_i}$ ▷ equation (16)
Update $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon}$ ▷ equation (17)
Update $G_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon}$ ▷ equation (18)

Outputs: $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}}$, $\{\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\zeta_i} : i = 1,\ldots,n\}$
$\{\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon}, G_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon}\}$

3.3 Functional Principal Component 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

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\boldsymbol{\nu}} \longleftarrow \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\Sigma}_\nu^{-1})\boldsymbol{\mu}_\nu \\ -\frac{1}{2}\,\text{vec}\left\{\mathbb{E}_q(\boldsymbol{\Sigma}_\nu^{-1})\right\} \end{bmatrix}. \tag{19}$$

where $\boldsymbol{\mu}_\nu \equiv (\boldsymbol{\mu}_\mu^\mathsf{T}, \boldsymbol{\mu}_{\psi_1}^\mathsf{T},\ldots,\boldsymbol{\mu}_{\psi_L}^\mathsf{T})^\mathsf{T}$ and $\boldsymbol{\Sigma}_\nu \equiv \text{blockdiag}(\boldsymbol{\Sigma}_\mu,\boldsymbol{\Sigma}_{\psi_1},\ldots,\boldsymbol{\Sigma}_{\psi_L})$. The natural pa-

rameter vector in (19) is within the vec-based representation of the message to $\boldsymbol{\nu}$, as opposed to a storage-economical vech-based representation, for the same reasons that are outlined in the discussion following (15).

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

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\rightarrow\sigma_\mu^2} \longleftarrow \begin{bmatrix} -\frac{K}{2} \\ -\frac{1}{2}\,\mathbb{E}_q(\boldsymbol{u}_\mu^\mathsf{T}\boldsymbol{u}_\mu) \end{bmatrix}. \tag{20}$$

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

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\rightarrow\sigma_{\psi_l}^2} \longleftarrow \begin{bmatrix} -\frac{K}{2} \\ -\frac{1}{2}\,\mathbb{E}_q(\boldsymbol{u}_{\psi_l}^\mathsf{T}\boldsymbol{u}_{\psi_l}) \end{bmatrix}. \tag{21}$$

Finally, recall the discussion following (17). Each of the messages to the variance parameters $\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2$ must be paired with a graph message. For the same reasons that were used to justify the graphical message in (18), the graph messages received by $\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2$ are, respectively,

$$G_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\rightarrow\sigma_\mu^2} \longleftarrow G_{\text{full}} \quad \text{and} \quad G_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\rightarrow\sigma_{\psi_l}^2} \longleftarrow G_{\text{full}}, \quad l=1,\ldots,L. \tag{22}$$

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

Algorithm 2 Pseudocode for the functional principal component Gaussian penalization fragment.

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

Updates:

Update posterior expectations. ▷ see Appendix E
Update $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_{\mu},\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\boldsymbol{\nu}}$ ▷ equation (19)
Update $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_{\mu},\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\mu}}$ ▷ equation (20)
Update $G_{p(\boldsymbol{\nu}|\sigma^2_{\mu},\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\mu}}$ ▷ equation (22)
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 (21)
    Update $G_{p(\boldsymbol{\nu}|\sigma^2_{\mu},\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\psi_l}}$ ▷ equation (22)

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

4 Post-VMP Steps

The FPCA model for curve estimation (2), which has its genesis in the Karhunen-Loève decomposition (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 curve estimation is still valid without these constraints, interpretation of the analysis is more straightforward with orthogonal eigenfunctions. Furthermore, the eigenfunctions are not guaranteed to be normalized. In the following sections, we outline a sequence of post-VMP steps that aid inference and interpretability for variational Bayes-based FPCA.

4.1 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 (11). 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 $\mathbb{C}\text{ov}_q(\boldsymbol{\nu})$ for $q^*(\boldsymbol{\nu})$ can be computed from (13), and the corresponding parameters $\mathbb{E}_q(\boldsymbol{\zeta}_i)$ and $\mathbb{C}\text{ov}_q(\boldsymbol{\zeta}_i)$ for $q^*(\boldsymbol{\zeta}_i)$, $i = 1, \ldots, n$, can be computed from (14). Note that we partition $\mathbb{E}_q(\boldsymbol{\nu})$ as $\mathbb{E}_q(\boldsymbol{\nu}) = \{\mathbb{E}_q(\boldsymbol{\nu}_\mu)^\intercal, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_1})^\intercal, \ldots, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_L})^\intercal\}^\intercal$.

4.2 Establishing 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, Härdle, & Kneip, 2009).

Now, we outline the steps to construct orthogonal functional principal component eigenfunctions and uncorrelated scores. The existence and uniqueness of the eigenfunctions 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 (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. Establish the posterior estimates of the mean function $\mathbb{E}_q\{\mu(\boldsymbol{t}_g)\} = \boldsymbol{C}_g\,\mathbb{E}_q(\boldsymbol{\nu}_\mu)$ and the functional principal components eigenfunctions $\mathbb{E}_q\{\psi_l(\boldsymbol{t}_g)\} = \boldsymbol{C}_g\,\mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})$, $l = 1, \ldots, L$. 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 $\boldsymbol{\Psi}$.

Next, define $\boldsymbol{\Xi} \equiv [\ \mathbb{E}_q(\zeta_1) \quad \cdots \quad \mathbb{E}_q(\zeta_n)\ ]^\intercal$. Set $\boldsymbol{m}_\zeta$ to be the $L \times 1$ sample mean vector of the column vectors of $\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal \boldsymbol{\Xi}^\intercal$, and set

$$\widehat{\mu}(\boldsymbol{t}_g) \equiv \mathbb{E}_q\{\mu(\boldsymbol{t}_g)\} + \boldsymbol{U}_\psi \boldsymbol{m}_\zeta. \tag{23}$$

Then set $\boldsymbol{C}_\zeta$ to be the $L \times L$ sample covariance matrix of the column vectors of $\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal \boldsymbol{\Xi}^\intercal - \boldsymbol{m}_\zeta \mathbf{1}_n^\intercal$ 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

$$\widetilde{\boldsymbol{\Psi}} \equiv \boldsymbol{U}_\psi \boldsymbol{Q} \boldsymbol{\Lambda}^{1/2} \quad \text{and} \quad \widetilde{\boldsymbol{\Xi}} \equiv (\boldsymbol{\Xi} \boldsymbol{V}_\psi \boldsymbol{D}_\psi - \mathbf{1}_n \boldsymbol{m}_\zeta^\intercal) \boldsymbol{Q} \boldsymbol{\Lambda}^{-1/2}. \tag{24}$$

Notice that $\widetilde{\boldsymbol{\Psi}}$ is an $n_g \times L$ matrix and $\widetilde{\boldsymbol{\Xi}}$ is an $n \times L$ matrix. Next, partition these matrices such that the lth column of $\widetilde{\boldsymbol{\Psi}}$ is $\widetilde{\psi}_l(\boldsymbol{t}_g)$ and the ith row of $\widetilde{\boldsymbol{\Xi}}$ is $(\widetilde{\zeta}_{11}, \ldots, \widetilde{\zeta}_{1L})$. The columns of $\widetilde{\boldsymbol{\Psi}}$ 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 $||\widetilde{\psi}_l||$, $l = 1, \ldots, L$, through numerical integration. This allows us to establish estimates of the orthonormal functions $\psi_1^*, \ldots, \psi_L^*$ in (2) over the vector $\boldsymbol{t}_g$ with

$$\widehat{\psi}_l(\boldsymbol{t}_g) \equiv \frac{\widetilde{\psi}_l(\boldsymbol{t}_g)}{||\widetilde{\psi}_l||}, \quad l = 1, \ldots, L, \tag{25}$$

as well as estimates of the scores with $\widehat{\zeta}_{il} \equiv ||\widetilde{\psi}_l|| \widetilde{\zeta}_{il}$. Lemma 4.1 outlines the construction of posterior curve estimation for the response vectors $y_1(\boldsymbol{t}_g), \ldots, y_n(\boldsymbol{t}_g)$. Proposition 4.2

shows that the form of the predicted response vectors in Lemma 4.1 is a vector 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 4.1. *The posterior estimate for the response vector $y_i(\boldsymbol{t}_g)$ is given by*

$$\widehat{y}_i(\boldsymbol{t}_g) = \widehat{\mu}(\boldsymbol{t}_g) + \sum_{l=1}^{L} \widehat{\zeta}_{il} \widehat{\psi}_l(\boldsymbol{t}_g), \quad i = 1, \ldots, n. \tag{26}$$

Remark. The posterior estimates $\widehat{y}_1(\boldsymbol{t}_g), \ldots, \widehat{y}_n(\boldsymbol{t}_g)$ in (26) are the same as those prior to the post-processing steps outlined above. That is, $\widehat{y}_i(\boldsymbol{t}_g) = \boldsymbol{C}_g \mathbb{E}_q(\boldsymbol{\nu}_\mu) + \sum_{l=1}^{L} \mathbb{E}_q(\zeta_{il}) \boldsymbol{C}_g \mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})$, where $\mathbb{E}_q(\boldsymbol{\nu}_\mu)$ is the posterior estimate of $\boldsymbol{\nu}_\mu$ from the VMP algorithm and similarly for $\mathbb{E}_q(\zeta_{il})$ and $\mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})$. In summary, the post processing steps simply realign the mean function, orthogonalize and normalize the eigenfunctions and uncorrelate the scores, but do not affect the fits to the observed data.

Proposition 4.2. *The vectors $\widehat{\boldsymbol{\zeta}}_1, \ldots, \widehat{\boldsymbol{\zeta}}_N$ are independent and satisfy $\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\zeta}}_i = \mathbf{0}$ and $\frac{1}{n-1}\sum_{i=1}^{n} \widehat{\boldsymbol{\zeta}}_i \widehat{\boldsymbol{\zeta}}_i^\intercal = \mathrm{diag}(||\widetilde{\psi}_1||^2, \ldots, ||\widetilde{\psi}_L||^2)$. Furthermore, the vectors $\widehat{\psi}_1(\boldsymbol{t}_g), \ldots, \widehat{\psi}_L(\boldsymbol{t}_g)$ are eigenvectors of the sample covariance matrix of $\widehat{y}_1(\boldsymbol{t}_g), \ldots, \widehat{y}_n(\boldsymbol{t}_g)$.*

Remark. Proposition 4.2 shows that the sample properties of the posterior estimates for the scores obey the assumptions of the scores in the Karhunen-Loève decomposition in (1). Furthermore, the vectors $\widehat{\psi}_1(\boldsymbol{t}_g), \ldots, \widehat{\psi}_L(\boldsymbol{t}_g)$ respect the orthogonality conditions in ℓ^2. Therefore, (26) may be interpreted as a vector version of the truncated Karhunen-Loève decomposition. As a consequence, the numerical estimates of $||\widetilde{\psi}_l||^2$, $l = 1, \ldots, L$ are the posterior estimates of the eigenvalues of the covariance operator Σ (see the first paragraph of Section 2).

The proof of Lemma 4.1 is presented in Appendix B, and the proof of Proposition 4.2 is presented in Appendix C.

5 Simulations

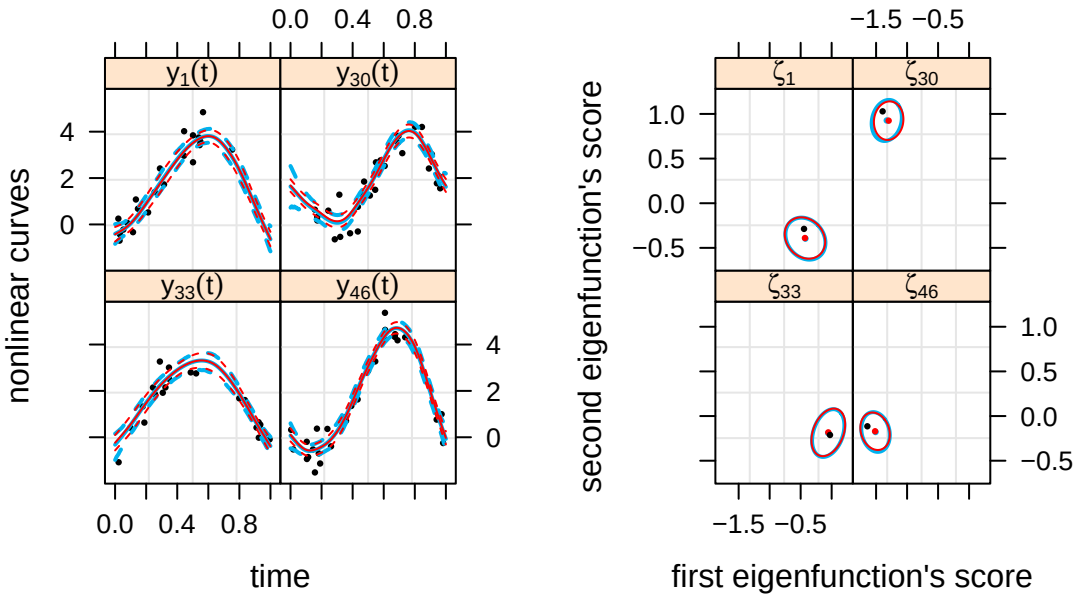
We illustrate the use of Algorithms 1 and 2 through a series of simulations of model (6). Pseudocode for the VMP algorithm is provided in Algorithm 3.

Algorithm 3 Generic VMP algorithm for the Gaussian response FPCA model (6) with mean field restriction (9).

Inputs: All hyperparameters and observed data
Initialize: All factor to stochastic node messages. ▷ Wand (2017, Section 2.5)
Updates:
while $\log \underline{p}(\boldsymbol{y};q)$ has not converged **do**
  Update all stochastic node to factor messages. ▷ Wand (2017, Section 2.5)
  Update the fragment for $p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)$ ▷ Algorithm 1
  Update the fragment for $p(\sigma^2_\varepsilon|a_\varepsilon)$ ▷ Maestrini and Wand (2020, Algorithm 2)
  Update the fragment for $p(a_\varepsilon)$ ▷ Maestrini and Wand (2020, Algorithm 1)
  Update the fragment for $p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})$ ▷ Algorithm 2
  for $i = 1,\ldots,n$ **do**
    Update the fragment for $p(\boldsymbol{\zeta}_i)$ ▷ Wand (2017, Section 4.1.1)
  Update the fragment for $p(\sigma^2_\mu|a_\mu)$ ▷ Maestrini and Wand (2020, Algorithm 2)
  Update the fragment for $p(a_\mu)$ ▷ Maestrini and Wand (2020, Algorithm 1)
  for $i = 1,\ldots,n$ **do**
    Update the fragment for $p(\sigma^2_{\psi_l}|a_{\psi_l})$ ▷ Maestrini and Wand (2020, Algorithm 2)
    Update the fragment for $p(a_{\psi_l})$ ▷ Maestrini and Wand (2020, Algorithm 1)
  Rotate, translate and re-scale $\boldsymbol{\Psi}$ and $\boldsymbol{\Xi}$. ▷ Section 4.2
Outputs: $\widehat{\mu}(\boldsymbol{t}_g)$, $\widehat{\psi}_1(\boldsymbol{t}_g),\ldots,\widehat{\psi}_L(\boldsymbol{t}_g)$ and $\widehat{\boldsymbol{\zeta}}_1,\ldots,\widehat{\boldsymbol{\zeta}}_n$.

5.1 Accuracy Assessment

For model (6), we simulated 50 response curves with the number of observations T_i for the ith curve sampled uniformly over $\{20,21,\ldots,30\}$. Furthermore, the time observations within the ith curve $\{t_{i1},\ldots,t_{iT_i}\}$ were sampled uniformly over the interval $(0,1)$, while the residual variance σ^2_ε was set to 1. The mean function was $\mu(t) = 3\sin(\pi t)$ and the eigenfunctions were $\psi_1(t) = \sqrt{2}\sin(2\pi t)$ and $\psi_2(t) = \sqrt{2}\cos(2\pi t)$. Each vector of principal component scores were simulated according to $\boldsymbol{\zeta}_i \equiv (\zeta_{i1},\zeta_{i2})^\mathsf{T} \overset{\text{ind.}}{\sim} \mathrm{N}\{(0,0)^\mathsf{T},\mathrm{diag}(1,0.25)\}$.



(a) (b)

Figure 2: The results from one simulation of the Gaussian response FPCA model in (6). The simulation parameters are outlined in Section 5. In (a), the simulated data are shown in black, while the VMP-based variational Bayes posterior estimates are presented in red and the corresponding MCMC estimates are shown in blue. In each panel, the solid lines represent the posterior mean, while the dashed line represents the 95% pointwise credible sets for the mean. In (b), we present the vector of scores for each of the randomly selected response curves, shown in black, as well as the VMP-based variational Bayes posterior estimates, shown in red, and the MCMC-based posterior estimates, shown in blue. The red and blue dots represent the VMP-based variational Bayes posterior means and the MCMC-based posterior means, respectively. The ellipses represent the 95% credible contours.

Nonparameteric regression with O'Sullivan penalized splines for the nonlinear curves was performed with $K = 10$. Finally, the simulations were conducted by setting $L = 3$, rather than 2 (the number of eigenfunctions), to assess the flexibility of the VMP algorithm under slight model misspecification.

The results from the simulation are presented in Figure 2, where a random sample of four of the functional responses are selected for visual clarity. In addition, we have included the results from an MCMC treatment of model (6) in blue for comparison with the VMP-

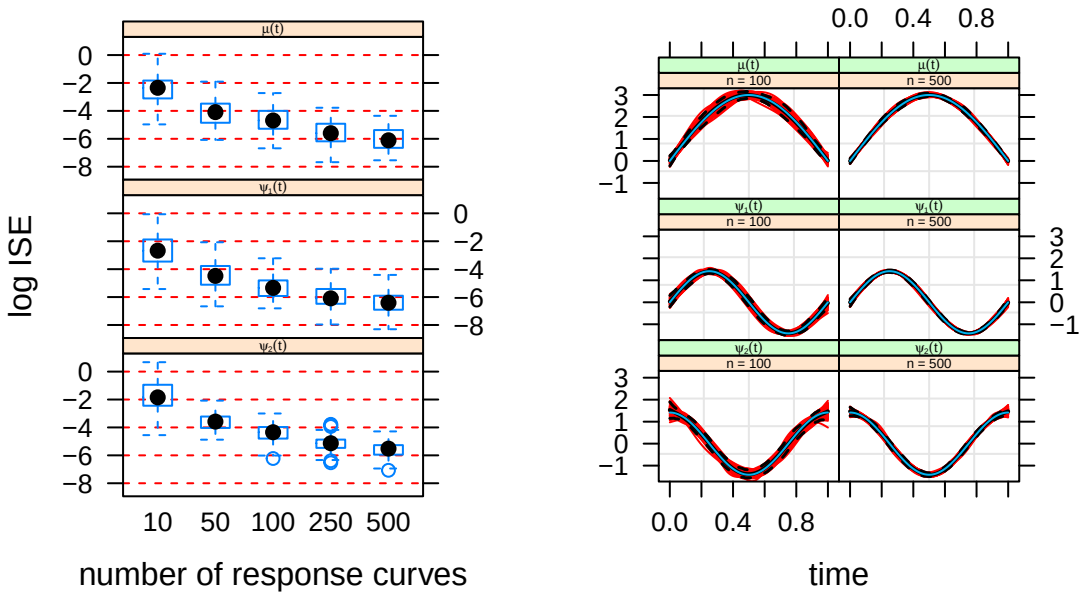
based variational Bayes fits in red. MCMC simulations were conducted through Rstan, the R (R Core Team, 2020) interface to the probabilistic programming language Stan (Stan Development Team, 2020). The variational Bayes fits have good agreement with their MCMC counterparts, as well as the simulated data. In particular, the post-VMP procedures that are outlined in Section 4 neatly complement the standard VMP algorithm.

We then incorporated five settings for the number of response curves: $n \in \{10, 50, 100, 250, 500\}$. For each of these settings, we conducted 100 simulations of model (6) with the aim of analysing the error of the posterior mean estimates of the mean curve and the functional principal component eigenfunctions. The error of each simulation was determined via the integrated squared error:

$$\text{ISE}(f, \widehat{f}) = \int_0^1 \left| f(x) - \widehat{f}(x) \right|^2 dx, \tag{27}$$

where, in our simulations, $f(\cdot)$ represents the true function that generated the data, while $\widehat{f}(\cdot)$ represents the VMP-based variational Bayes posterior mean curve.

The box plots for the logarithm of the integrated squared error values in Figure 3 (a) reflect the excellent results for the settings where $n = 50,\ 100,\ 250$ and 500. Overall, the results for the setting where $n = 10$ are good, however, there are a few simulations where the posterior estimates of the second functional principal component eigenfunction $\psi_2(\cdot)$ are poor. This is to be expected because the scores associated with this eigenfunction were generated from a $\text{N}(0, 0.25)$ distribution reflecting its weaker contribution to the data generation process. Also, as expected, the ISE for all curves tend to decline with increasing n. In Figure 3 (b), we present all of the simulated posterior mean curves of the mean function and the functional principal component eigenfunctions, for the case where $n = 100$ and $n = 500$, which are overlaid with the true functions in blue. These plots demonstrate the strength of the VMP algorithm for estimating the underlying curves that generate an



(a) (b)

Figure 3: The results from a simulation study of the Gaussian response FPCA model in (6). The simulation parameters are outlined in Section 5.1. The box plots in (a) are a summary of the logarithm of the integrated squared error values in (27) for 100 simulations of each of the settings $n \in \{10, 50, 100, 250, 500\}$. In (b), we present the results for the mean function and the eigenfunctions when $n = 100$ (left column) and $n = 500$ (right column). The true functions are shown in blue in each panel, and the VMP-based posterior mean curve for each of the simulations is presented in red. In addition, we have included the pointwise mean curve and the pointwise 95% confidence intervals resulting from the MCMC posterior estimates of all of the generated datasets in black. Note that, in each panel, the pointwise MCMC mean curve overlaps very tightly with the true curve, making it difficult to see.

observed set of functional data. Furthermore, the variability in the curve estimates is drastically reduced with increasing n. In addition to each of the VMP posterior estimates, we have included the pointwise mean curve and the pointwise 95% confidence interval for the MCMC simulations of all of the generated data sets. Evidently, there is strong agreement between the VMP simulations and the MCMC simulations.

5.2 Computational Speed Comparisons

In the previous section, we saw that the mean field product restriction in (9) does not compromise the accuracy of variational Bayesian inference for FPCA. However, the major advantage offered by variational Bayesian inference via VMP is fast approximate inference in comparison to MCMC simulations. Several published articles have addressed the computational speed gains from using variational Bayesian inference. Faes, Ormerod, and Wand (2011) presented speed gains for parametric and nonparametric regression with missing data, Luts and Wand (2015) presented timing comparisons for semiparametric regression models with count responses, and Lee and Wand (2016) and Nolan, Menictas, and Wand (2020) established speed gains for multilevel data models with streamlined matrix algebraic results. In all cases, the variational Bayesian inference algorithms had strong accuracy in comparison to MCMC simulations and were far superior in computational speed.

In Table 1, we present a similar set of results for the computational speed of VMP and MCMC for model (6). The simulations were identical to those that were used to generate the results in Figure 3, where there were 100 simulations over five settings for the number of response curves $n \in \{10, 50, 100, 250, 500\}$. In addition, the simulations were performed on a laptop computer with 8 GB of random access memory and a 1.6 GHz processor. In Table 1, we present the median elapsed computing time (in seconds), with the first quartile and the third quartile shown in brackets. Notice that most of the VMP simulations are completed within 1 minute, whereas the elapsed computing time for the MCMC simulations tends to vary from approximately 1 minute, for $n = 10$, to over an hour, for $n = 500$. The most impressive results are in the fourth column, where the median VMP simulation is 19.6 times faster than the median MCMC simulation 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 $n = 500$.

Table 1: Median (first quartile, third quartile) elapsed computing time in seconds for VMP and MCMC with $n \in (10, 50, 100, 250, 500)$. The fourth column presents the ratio of the median elapsed time for MCMC to the median elapsed time for VMP.

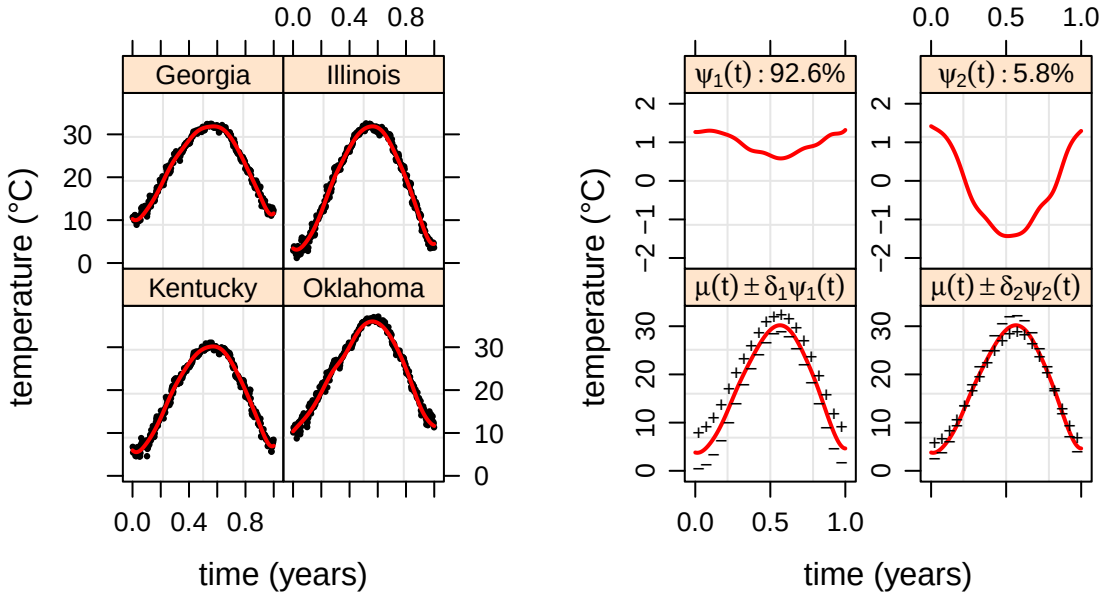
n	VMP	MCMC	MCMC/VMP
10	2.1 (1.3, 2.8)	41.2 (37.3, 45.6)	19.6
50	5.5 (3.4, 9.9)	188.4 (178.8, 213.9)	34.3
100	11.8 (7.0, 18.2)	446.7 (415.1, 475.7)	37.9
250	33.1 (18.1, 48.6)	1620.8 (1446.6, 1864.7)	49.0
500	58.0 (32.1, 91.0)	3471.2 (2832.9, 4497.8)	59.8

6 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 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 state or the federal district of the site. Our analysis focused on maximum daily temperature that was averaged over the 25 years of available data.

From this package, we collected full data sets (data available for every day of the year) from 2837 weather stations, where 49 states and federal districts were represented. For each state or federal district, we took a random sample of 3 of the available sites. In cases where there were less than 3 sites available (Rhode Island and District of Columbia), we used all available sites. This resulted in 145 sites used in our application, with 365 observations for each site.

Chapter 8 of Ramsay and Silverman (2005) conducts a similar analysis of Canadian



(a) (b)

Figure 4: Application of the VMP algorithm for FPCA to the United States temperature data. The fits in (a) are for four randomly selected weather stations in the dataset. The plots in (b) present the pointwise posterior mean estimates of the eigenfunctions (top panel) and show the estimated mean function with perturbations from each eigenfunction (bottom panel): $\widehat{\mu}(t) \pm \delta_l \widehat{\psi}_l(t)$, $l = 1,2$.

temperature data from various weather stations. In their application, they uncovered four functional principal component eigenfunctions. Similarly, we conducted VMP simulations with $L = 4$. The results are presented in Figure 4. In Figure 4 (a), we display the results of four randomly selected weather stations, from four different states. There is relatively small residual variability in the observed dataset because we are using long-term averages. As a consequence, the pointwise 95% credible sets would not be visible in the plots, so we have only included the pointwise variational Bayesian posterior means. In Figure 4 (b), we present the pointwise posterior estimates for the first two eigenfunction (top panel) and the the effect of perturbing the estimated mean function with each eigenfunction (bottom panel): $\widehat{\mu}(t) \pm \delta_l \widehat{\psi}_l(t)$, $l = 1,2$. The plus (minus) signs indicate the shift that each eigenfunction makes to the mean function with a positive (negative) perturbation. In addi-

tion, the value of δ_l was simply selected such that the effect of the perturbation would be visibly apparent. Note that the top and bottom panels in Figure 4 (b) should be analysed concurrently when determining the effect of each eigenfunction.

The bottom panel for the first eigenfunction (which accounts for 92.6% of the total variation) shows that it is a mean shift that perturbs the mean function in the positive (negative) direction when it is added (subtracted). The top panel shows that this effect is stronger in the Winter months than the Summer months, indicating that US temperature is most variable in Winter. Similar analysis of the second eigenfunction (which accounts for 5.8% of the total variation) shows that it represents uniformity in the measured temperatures. It perturbs the mean function in the negative (positive) direction in the Summer (Winter) months when it is added. 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. The third and fourth eigenfunctions were harder to interpret given their weak contributions to the total variation and were omitted from the analysis. The scores associated with the first two eigenfunctions for the displayed weather stations are (4.95, 0.35) for the weather station in Georgia, (0.79, -1.11) for the weather station in Illinois, (1.55, 0.16) for the weather station in Kentucky and (6.63, -0.86) for the weather station in Oklahoma. The scores for the first eigenfunction indicate that higher than average temperatures are expected in all four states, with this effect most pronounced in Georgia and Oklahoma. After accounting for displacement from the mean temperature function, the scores for the second eigenfunction indicate that the greatest variability between Summer and Winter months can be found in Illinois and Oklahoma, whereas Georgia and Kentucky appear to be more uniform.

7 Closing Remarks

We have provided a comprehensive overview of Bayesian FPCA 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 two new fragments for VMP:

1. the functional principal component Gaussian likelihood fragment (Algorithm 1);

2. and the functional principal component Gaussian penalization fragment (Algorithm 2).

These are directly compatible with the fragment-based computational constructions of VMP outlined in Wand (2017). This is, to our knowledge, the first VMP construction of a Bayesian FPCA model. In addition, we have outlined a sequence of post-VMP steps that are necessary for producing orthonormal functional principal component eigenfunctions and uncorrelated scores.

Simulations were conducted to assess the speed and accuracy of the VMP simulations against MCMC counterparts. The approximate variational posterior density functions were in good agreement with the MCMC estimations, and the VMP algorithm was approximately 20 to 60 times faster than the MCMC algorithm depending on the number of response curves. An application to a large US temperature dataset showed that the VMP-based FPCA algorithm can be used for strong inference in big data applications.

This study could be extended to other functional data models, such as function on scalar or vector regression models, that are yet to be treated under a VMP-based mean field variational Bayes approach. In addition, extending the likelihood specification to generalized outcomes would also satisfy a popular area of research in functional data analysis.

Funding

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.

References

Benko, M., Härdle, W., & Kneip, A. (2009). Common functional principal components. *The Annals of Statistics*, *37*, 1–34.

Bishop, C. M. (1999). Variational pincipal components. In *Proceedings of the ninth international conference on artificial neural networks.* Institute of Electrical and Electronics Engineers.

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

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*, 859–877.

Chamberlain, S., Anderson, B., Salmon, M., Erickson, A., Potter, N., Stachelek, J., ... rOpenSci (2021). *'NOAA' weather data from r.* Retrieved from `https://docs.ropensci.org/rnoaa/` (R package version 1.3.0)

Di, C. Z., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics*, *3*, 458–488.

Durbán, M., Harezlak, J., Wand, M. P., & Carroll, R. J. (2005). Simple fitting of subject specific curves for longitudinal data. *Statistics in Medicine*, *24*, 1153–1167.

Faes, C., Ormerod, J. T., & Wand, M. P. (2011). Variational Bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American*

Statistical Association, *106*, 959–971.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*, 515–534.

Gentle, J. E. (2007). *Matrix Algebra*. New York: Springer.

Goldsmith, J., Greven, S., & Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, *69*, 41–51.

Goldsmith, J., & Kitago, T. (2016). Assessing systematic effects of stroke on motorcontrol by using hierarchical function-on-scalar regression. *Journal of the Royal Statistical Society. Series C, Applied statistics*, *65*, 215–236.

Goldsmith, J., Zippunnikov, V., & Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, *71*, 344–353.

Greven, S., Crainiceanu, C., Caffo, B., & Reich, D. (2011). Longitudinal functional principal component analysis. In *Recent advances in functional data analysis and related topics* (pp. 149–154). Springer.

James, G. M., Hastie, T. J., & Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, *87*(3), 587–602.

Lee, C. Y. Y., & Wand, M. P. (2016). Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal*, *58*, 868–895.

Luts, J., & Wand, M. P. (2015). Variational inference for count response semiparametric regression. *Bayesian Analysis*, *10*, 991–1023.

Maestrini, L., & Wand, M. P. (2020). The Inverse G-Wishart distribution and variational message passing. *arXiv e-prints*, arXiv:2005.09876.

Menictas, M., & Wand, M. P. (2013). Variational inference for marginal longitudinal semiparametric regression. *Stat*, *2*, 61–71.

Minka, T. (2005). *Divergence measures and message passing* (Tech. Rep.). Cambridge, UK: Microsoft Research Ltd.

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

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*, 140–153.

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

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, *3*, 1193–1256.

Stan Development Team. (2020). *RStan: the R interface to Stan*. Retrieved from `http://mc-stan.org/` (R package version 2.21.2)

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, *3*, 611–622.

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

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.

Wand, M. P., & Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, *50*, 179–198.

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

Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, *6*, 661–694.

Xiao, L., Zipunnikov, V., Ruppert, D., & Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data. *Statistics and computing*, *26*(1-2), 409–421.

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