Bayesian Functional Principal Components Analysis via Variational Message Passing

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

[1]School of Operations Research and Information Engineering, Cornell University
[2]School of Mathematical and Physical Sciences, University of Technology Sydney
[3]Mailman School of Public Health, Columbia University
[4]School of Statistical Science, Cornell University

February 22, 2021

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 maximise a mean square criterion. In the functional setting, response curves are interpreted as independent realisations of an underlying stochastic process. A covariance operator and its basis functions play the analogous role that the covariance matrix and its eigenvectors play in the multivariate data setting. By determining the basis functions that have the strongest contributions to the covariance function, one can reduce the dimensionality of the entire data set by approximating each curve as a linear combination of the reduced set of basis functions.

There are key issues that arise in the functional setting, which are not present in the multivariate setting. 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 semiparametric regression to infer the

*Corresponding author: thn22@cornell.edu

behaviour of the original curves and employ an appropriate orthogonalisation step to ensure that the basis functions are orthogonal.

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 eigenfuncion basis 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 estimated eigenbasis to estimate scores, meaning that inference on individual curve estimates can be poor.

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 EM 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, 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 determined, meaning that complex bivariate smoothing is not required. Indeed, the eigenfunctions and eigenvalues are computed directly as part of a Bayesian hierarchical model. Further, 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 eigenfunctions in a Bayesian model is straightforward. For these reason, we pursue a Bayesian approach to FPCA.

Although there have been numerous contributions to Bayesian implementations of FPCA, we argue that there are additional factors that should be considered. 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) 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 semiparametric regression 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. We provide an introduction to variational Bayesian inference in Section 2, with an overview of VMP in Section 2.2. Section 3 introduces the Bayesian hierarchical model for FPCA and its extensions under a VMP formulation. In Section 4, we outline the post-VMP steps that are required for producing orthogonal eigenfunctions. Simulations, including speed and accuracy comparisons with MCMC algorithms, are presented in Section 5, and an application to United States weather data is provided in Section 6.

2 Variational Bayesian Inference

The overarching aim of this article is the identification and derivation of fragments that are necessary for VMP implementations of FPCA. VMP is an approximate Bayesian inference construction that is derived from MFVB approaches. In this section, we provide a brief introduction to the MFVB and VMP frameworks. For an in-depth explanation of MFVB, we refer the reader to Ormerod and Wand (2010) and Blei, Kucukelbir, and McAuliffe (2017); for a comprehensive review of VMP, we refer the reader to Minka (2005) and Wand (2017).

Variational Bayesian inference is based on the notion of minimal Kullback-Leibler divergence to approximate a posterior density function. For arbitrary density functions p_1 and p_2 on $\mathbb{R}^d$, the Kullback-Leibler divergence of p_1 from p_2 is

$$D_{\mathrm{KL}}(p_1, p_2) \equiv \int_{\mathbb{R}^d} \log\left\{\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right\} p_1(\boldsymbol{x}) d\boldsymbol{x}.$$

Note that

$$D_{\mathrm{KL}}(p_1, p_2) \geq 0. \tag{1}$$

Consider a generic Bayesian model with observed data vector $\boldsymbol{y}$ and parameter vector $\boldsymbol{\theta} \in \Theta$, where Θ is a parameter space. We make the assumption that $\boldsymbol{y}$ and $\boldsymbol{\theta}$ are continuous random variables with density functions $p(\boldsymbol{y})$ and $p(\boldsymbol{\theta})$. For the case where some components are discrete, a similar treatment applies with summations replacing integrals. Next, let $q(\boldsymbol{\theta})$ represent an arbitrary density function over the parameter space Θ. The essence of variational Bayesian inference is to restrict q to a class of density functions Q and use the optimal q-density function, defined by

$$q^*(\boldsymbol{\theta}) \equiv \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\ D_{\mathrm{KL}}\{q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\boldsymbol{y})\}, \tag{2}$$

as an approximation to the true posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$.

Simple algebraic arguments (e.g. Ormerod & Wand, 2010, Section 2.1) show that the marginal log-likelihood satisfies:

$$\log p(\boldsymbol{y}) = D_{\mathrm{KL}}\{q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\boldsymbol{y})\} + \log \underline{p}(\boldsymbol{y}; q),$$

where

$$\underline{p}(\boldsymbol{y}; q) \equiv \exp\left[\int \log\left\{\frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\} q(\boldsymbol{\theta}) d\boldsymbol{\theta}\right].$$

From the non-negativity condition of (1), we have

$$\underline{p}(\boldsymbol{y}; q) \leq p(\boldsymbol{y})$$

showing that $\underline{p}(\boldsymbol{y}; q)$ is a lower-bound on the marginal likelihood. This leads to an equivalent form for the optimisation problem in (2):

$$q^*(\boldsymbol{\theta}) \equiv \underset{q \in \mathcal{Q}}{\operatorname{argmax}}\{\log \underline{p}(\boldsymbol{y}; q)\}. \tag{3}$$

As stated in (Rohde & Wand, 2016), this alternate expression has the advantage of representing the optimal q-density function as maximising the lower-bound on the marginal log-likelihood. For the remainder of this article, we will address variational Bayesian inference with (3), rather than (2).

2.1 Mean Field Variational Bayes

MFVB is a class of variational Bayesian inference methods that uses a product density (or mean field) restriction in the optimal q-density function. The mean field approximation, which has its roots in statistical physics (Parisi, 1988), imposes the factorization

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{N} q(\boldsymbol{\theta}_i), \tag{4}$$

for all $q \in Q$, where $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$ is some partition of $\boldsymbol{\theta}$. The optimal q-density functions that satisfy (3) are given by (e.g. Minka, 2005; Ormerod & Wand, 2010)

$$q_i^*(\boldsymbol{\theta}_i) = \frac{\exp\{\mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_i)} \log p(\boldsymbol{y}, \boldsymbol{\theta})\}}{\int \exp\{\mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_i)} \log p(\boldsymbol{y}, \boldsymbol{\theta})\} d\boldsymbol{\theta}_i}, \quad \text{for } i = 1, \ldots, N,$$

where $\mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_i)}$ denotes expectation with respect to the optimal posterior density functions of all elements in the partition of $\boldsymbol{\theta}$, defined by (4), except for the optimal posterior density function of $\boldsymbol{\theta}_i$. The standard MFVB algorithm is presented as Algorithm 1 in Ormerod and Wand (2010).

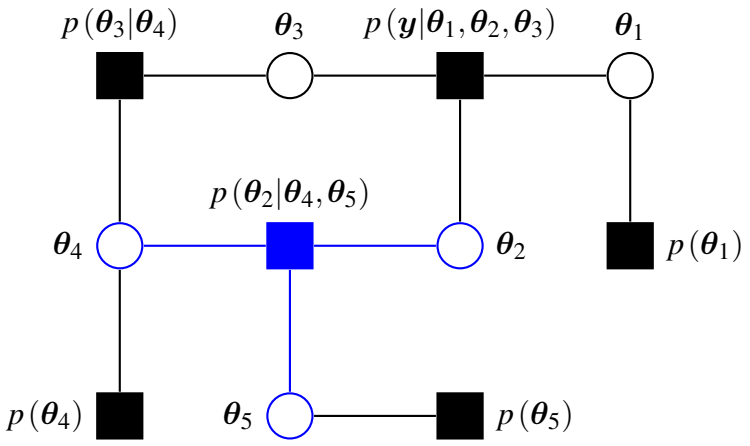


Figure 1: A factor graph representation of the Bayesian model described by (5). As an example, the factor graph fragment for $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)$ is highlighted in blue.

2.2 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. Here, we give a brief description of the foundations of VMP. For a thorough exposition of the VMP framework, we refer the reader to Wand (2017).

As a specific example, consider a generic Bayesian model with observed data vector $\boldsymbol{y}$ and parameter vectors $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_5$. Suppose that the joint density function factorises according to

$$p(\boldsymbol{y},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_5) = p(\boldsymbol{y}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3)p(\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_4)p(\boldsymbol{\theta}_4)p(\boldsymbol{\theta}_5). \tag{5}$$

A factor graph representation of the Bayesian model expressed in (5) is presented in Figure 1. The square nodes are the factors, which represent the distributional specifications of the model, and the circular nodes are called stochastic nodes, which represent the random variables of the model. Furthermore, the graph is bipartite, meaning that stochastic nodes can only share an edge with a factor and vice versa. Additionally, notice that the edges of the factor graph respect the distributional dependencies of (5). For instance, the factor for $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)$ shares edges with the stochastic nodes for $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_4$ and $\boldsymbol{\theta}_5$ only.

The VMP construction of variational Bayesian inference relies on messages passed between factors and stochastic nodes. Consider the factor for $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)$ and the messages that it will pass to its neighboring stochastic nodes $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_4$ and $\boldsymbol{\theta}_5$. The messages passed from this factor take the form

$$m_{p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow \frac{\mathbb{E}_{q(\boldsymbol{\theta}\setminus\boldsymbol{\theta}_i)}\log p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)}{\int \mathbb{E}_{q(\boldsymbol{\theta}\setminus\boldsymbol{\theta}_i)}\log p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)d\boldsymbol{\theta}_i}, \quad \text{for } i=2,4,5.$$

Next, consider the factor $\boldsymbol{\theta}_4$, which receives messages from the factors $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)$, $p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_4)$ and $p(\boldsymbol{\theta}_4)$. The messages that $\boldsymbol{\theta}_4$ returns to these factors are

$$
\begin{aligned}
m_{\boldsymbol{\theta}_4 \to p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)}(\boldsymbol{\theta}_4) &\longleftarrow m_{p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_4)\to\boldsymbol{\theta}_4}(\boldsymbol{\theta}_4) m_{p(\boldsymbol{\theta}_4)\to\boldsymbol{\theta}_4}(\boldsymbol{\theta}_4) \\
m_{\boldsymbol{\theta}_4 \to p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_4)}(\boldsymbol{\theta}_4) &\longleftarrow m_{p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)\to\boldsymbol{\theta}_4}(\boldsymbol{\theta}_4) m_{p(\boldsymbol{\theta}_4)\to\boldsymbol{\theta}_4}(\boldsymbol{\theta}_4) \\
m_{\boldsymbol{\theta}_4 \to p(\boldsymbol{\theta}_4)}(\boldsymbol{\theta}_4) &\longleftarrow m_{p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)\to\boldsymbol{\theta}_4}(\boldsymbol{\theta}_4) m_{p(\boldsymbol{\theta}_3|\boldsymbol{\theta}_4)\to\boldsymbol{\theta}_4}(\boldsymbol{\theta}_4).
\end{aligned}
$$

Now, consider a general Bayesian model with M factors $p_1, \ldots, p_M$ and N parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. We define the set of stochastic nodes that are connected to the factor p_j as

$$
\text{neighbors}(j) \equiv \{i = 1, \ldots, N : \boldsymbol{\theta}_i \text{ is shares an edge with } p_j\}.
$$

Then, the message from factor p_j to the stochastic node $\boldsymbol{\theta}_i$ is

$$
m_{p_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow \frac{\mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_i)} \log p_j}{\int \mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_i)} \log p_j d\boldsymbol{\theta}_i}. \tag{6}
$$

The message from $\boldsymbol{\theta}_i$ to p_j is

$$
m_{\boldsymbol{\theta}_i \to p_j}(\boldsymbol{\theta}_i) \longleftarrow \prod_{j' \neq j : i \in \text{neighbors}(j')} m_{p_{j'} \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i). \tag{7}
$$

Upon convergence of the messages, the optimal q-density functions, which satisfy (3) take the form

$$
q^*(\boldsymbol{\theta}_i) \propto \prod_{j : i \in \text{neighbors}(j)} m_{p_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i), \quad i = 1, \ldots, N. \tag{8}
$$

2.2.1 Exponential Family Form

A key step in deriving and implementing VMP algorithms is the representation of probability density functions in exponential family form. In particular, the messages in (6) and (7) are typically in the exponential family of density functions with vector of sufficient statistic $\boldsymbol{T}(\boldsymbol{\theta}_i)$. We have

$$
m_{p_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \propto \exp\{\boldsymbol{T}(\boldsymbol{\theta}_i)^\intercal \boldsymbol{\eta}_{p_j \to \boldsymbol{\theta}_i}\}, \quad \text{and} \quad m_{\boldsymbol{\theta}_i \to p_j}(\boldsymbol{\theta}_i) \propto \exp\{\boldsymbol{T}(\boldsymbol{\theta}_i)^\intercal \boldsymbol{\eta}_{\boldsymbol{\theta}_i \to p_j}\},
$$

where $\boldsymbol{\eta}_{p_j \to \boldsymbol{\theta}_i}$ and $\boldsymbol{\eta}_{\boldsymbol{\theta}_i \to p_j}$ are the message natural parameter vectors. Wand (2017) explains how these natural parameter vectors play a central role in the messages that are passed within and between factor graph fragments. Furthermore, the natural parameter vectors for the optimal q-density functions in (8) take the form

$$
\boldsymbol{\eta}_{q^*(\boldsymbol{\theta}_i)} = \sum_{j : i \in \text{neighbors}(j)} \boldsymbol{\eta}_{p_j \to \boldsymbol{\theta}_i}, \quad i = 1, \ldots, N. \tag{9}
$$

In addition, we adopt the notation

$$
\boldsymbol{\eta}_{p_j \leftrightarrow \boldsymbol{\theta}_i} \equiv \boldsymbol{\eta}_{p_j \to \boldsymbol{\theta}_i} + \boldsymbol{\eta}_{\boldsymbol{\theta}_i \to p_j}. \tag{10}
$$

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

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 \\ -3 & 1 \end{bmatrix}.$$

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 $\boldsymbol{D}_d^+ \text{vec}(\boldsymbol{A}) = \text{vech}(\boldsymbol{A})$.

The Normal distribution is one of the most important distributions within the exponential family, and it plays a major role in VMP versions of variational Bayesian inference. Consider the $d \times 1$ Multivariate Normal random vector $\boldsymbol{x} \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The probability density function of $\boldsymbol{x}$ can be shown to satisfy

$$p(\boldsymbol{x}) = \exp\left\{ \boldsymbol{T}_{\text{vec}}(\boldsymbol{x})^\intercal \boldsymbol{\eta}_{\text{vec}} - A_{\text{vec}}(\boldsymbol{\eta}_{\text{vec}}) - \frac{d}{2}\log(2\pi) \right\} \tag{11}$$

where

$$\boldsymbol{T}_{\text{vec}}(\boldsymbol{x}) \equiv \begin{bmatrix} \boldsymbol{x} \\ \text{vec}(\boldsymbol{x}\boldsymbol{x}^\intercal) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\eta}_{\text{vec}} \equiv \begin{bmatrix} \boldsymbol{\eta}_{\text{vec},1} \\ \boldsymbol{\eta}_{\text{vec},2} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix}$$

are, respectively, the vector of sufficient statistics and the natural parameter vector. The function

$$A_{\text{vec}}(\boldsymbol{\eta}_{\text{vec}}) = -\frac{1}{4}\boldsymbol{\eta}_{\text{vec},1}^\intercal \left\{\text{vec}^{-1}(\boldsymbol{\eta}_{\text{vec},2})\right\}^{-1} \boldsymbol{\eta}_{\text{vec},1} - \frac{1}{2}\log\left|-2\,\text{vec}^{-1}(\boldsymbol{\eta}_{\text{vec},2})\right|$$

is the log-partition function. The inverse mapping of the natural parameter vector is (Wand, 2017, equation S.4)

$$\boldsymbol{\mu} = -\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{\eta}_{\text{vec},2})\right\}^{-1} \boldsymbol{\eta}_{\text{vec},1} \quad \text{and} \quad \boldsymbol{\Sigma} = -\frac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{\eta}_{\text{vec},2})\right\}^{-1}. \tag{12}$$

We will refer to the representation of the Multivariate Normal probability density function in (11) 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, respectively,

$$\boldsymbol{T}_{\text{vech}}(\boldsymbol{x}) \equiv \begin{bmatrix} \boldsymbol{x} \\ \text{vech}(\boldsymbol{x}\boldsymbol{x}^\intercal) \end{bmatrix}, \quad \boldsymbol{\eta}_{\text{vech}} \equiv \begin{bmatrix} \boldsymbol{\eta}_{\text{vech},1} \\ \boldsymbol{\eta}_{\text{vech},2} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{D}_d^\intercal \text{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix}$$

and

$$A_{\text{vech}}(\boldsymbol{\eta}_{\text{vech}}) = -\frac{1}{4}\boldsymbol{\eta}_{\text{vech},1}^{\intercal}\left\{\text{vec}^{-1}(\boldsymbol{D}_d^{+\intercal}\boldsymbol{\eta}_{\text{vech},2})\right\}^{-1}\boldsymbol{\eta}_{\text{vech},1} - \frac{1}{2}\log\left|-2\,\text{vec}^{-1}(\boldsymbol{D}_d^{+\intercal}\boldsymbol{\eta}_{\text{vech},2})\right|.$$

The inverse mapping of the natural parameter vector under the vech-based representation is

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

The other major distribution within the exponential family that is pivotal for this article is the Inverse $-\chi^2$ distribution. A random variable x has an Inverse $-\chi^2$ 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 $\Gamma(\cdot)$ is the gamma function defined by $\Gamma(z) \equiv \int_0^\infty u^{z-1}e^u du$. The exponential family representation of the Inverse $-\chi^2$ density function is

$$p(x) = \exp\{\boldsymbol{T}(x)^{\intercal}\boldsymbol{\eta} - A(\boldsymbol{\eta})\}\,\mathbb{I}(x>0),$$

where the vector of sufficient statistics, the natural parameter vector and the log-partition function are, respectively,

$$\boldsymbol{T}(x) \equiv \begin{bmatrix}\log(x)\\ 1/x\end{bmatrix} \quad \text{and} \quad \boldsymbol{\eta} \equiv \begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} \equiv \begin{bmatrix}-\frac{1}{2}(\xi+2)\\ -\frac{\lambda}{2}\end{bmatrix}$$

and

$$A(\boldsymbol{\eta}) \equiv \log\{\Gamma(\xi/2)\} - \frac{\xi}{2}\log(\lambda/2).$$

The inverse mapping of the natural parameter vector is (Maestrini & Wand, 2020, equation 8)

$$\xi = -2\eta_1 - 2 \quad \text{and} \quad \lambda = -2\eta_2.$$

The generalization of the Inverse $-\chi^2$ distribution is the Inverse G-Wishart distribution, written $\boldsymbol{X} \sim \text{Inverse G-Wishart}(G, \xi, \Lambda)$, for a symmetric and positive definite $d \times d$ matrix $\boldsymbol{X}$. The parameter G is an undirected graph with d nodes and edge set E with pairs of nodes connected by an edge. Furthermore, $\xi > 0$ and $\boldsymbol{\Lambda}$ is a symmetric and positive definite $d \times d$ matrix. We say that, the symmetric matrix $\boldsymbol{A}$ respects G if $\boldsymbol{A}_{ij} = 0$ for all $(i,j) \notin E$. Now, impose the additional constraint that $\boldsymbol{X}^{-1}$ respects G. Furthermore, let G_{full} represent a complete graph where each node is connected to every other node by an edge, and let G_{diag} represent a sparse graph where the edge set is empty. The justification for this notation is that a matrix with no zero constraints respects G_{full} and a diagonal matrix respects G_{diag}. The two graph constructions generate two definitions for the Inverse G-Wishart

distribution (Maestrini & Wand, 2020, Definition 3):

(a) If $G = G_{\text{full}}$ and ξ is restricted such that $\xi > 2d-2$ then we say that $\boldsymbol{X}$ has an Inverse G-Wishart distribution with graph G, shape parameter ξ and scale matrix $\boldsymbol{\Lambda}$ if and only if the non-zero values of the density function of $\boldsymbol{X}$ satisfy

$$p(\boldsymbol{X}) \propto |\boldsymbol{X}|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\right\}$$

(b) If $G = G_{\text{diag}}$, then we say the $\boldsymbol{X}$ has an Inverse G-Wishart distribution with graph G, shape parameter ξ and scale matrix $\boldsymbol{\Lambda}$ if and only if the non-zero values of the density function of $\boldsymbol{X}$ satisfy

$$p(\boldsymbol{X}) \propto |\boldsymbol{X}|^{-(\xi+2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{\Lambda}\boldsymbol{X}^{-1})\right\}.$$

The exponential family form of the Inverse G-Wishart distribution is not important for this article; we refer the reader to Maestrini and Wand (2020, Section 2.2). Instead, our interest lies in the role of the graphical parameter, which must be incorporated as an argument for variational message updates involving Inverse G-Wishart random matrices, including Inverse $-\chi^2$ random variables. We follow the advice in Maestrini and Wand (2020, Sections 5 & 6) for message passing updates involving graphical parameters.

2.2.2 Factor Graph Fragments

A factor graph fragment (or fragment, for short) is the computational unit of VMP. A fragment, as defined by Wand (2017), is a subgraph of a factor graph consisting of a single factor and each of its neighboring stochastic nodes. An example of a factor graph fragment is the fragment for $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)$ in Figure 1, which is highlighted in blue. The factor representing $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_4,\boldsymbol{\theta}_5)$, the neighboring stochastic nodes $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_4$ and $\boldsymbol{\theta}_5$ and the connecting edges are all part of the fragment. The identification and derivation of the algebraic computations in fragments is the key step for extending VMP to arbitrarily large Bayesian statistical models. The following is a list of some of the major contributions to fragment updates for VMP:

- Wand (2017) identified five fundamental fragments for Gaussian response semiparametric regression. This includes the Gaussian prior fragment, Inverse G-Wishart prior fragment and the iterated Inverse G-Wishart fragment, which are necessary for a VMP construction of Bayesian FPCA. The author also identified several other fragments for direct implementation in various extensions for semiparametric regression.

- Nolan and Wand (2017) developed fast, stable and accurate numerical integration techniques for the logistic likelihood fragment. This had been previously introduced in Wand (2017) using the variational lower bound of Jaakkola and Jordan (2000), however the performance of this approximation can be poor (Knowles & Minka, 2011). Instead, Nolan and Wand (2017) incorporated the normal scale mixture uniform approximation of the logistic function (Monahan & Stefanski, 1992) into the logistic likelihood fragment for highly accurate inference.

- Maestrini and Wand (2018) derived algorithmic updates for the skew t likelihood fragment with all skew t parameters inferred, rather than being held fixed.

- McLean and Wand (2019) built on previous VMP constructions by focusing on regression models where the response variable is modeled to have an elaborate distribution, such as Negative Binomial and t likelihoods.

- Nolan, Menictas, and Wand (2020) used a set of solutions to sparse multilevel matrix problems (Nolan & Wand, 2020) to streamline the computations of VMP for Gaussian response linear mixed models. This involved the introduction of four new fragments.

- Maestrini and Wand (2020) provide corrections for the matrix versions of the Inverse G-Wishart prior fragment and the iterated Inverse G-Wishart fragment from Wand (2017). The corrections are based on graph theoretic results that affect VMP treatments of covariance matrices. Although the scalar versions of these fragments from Wand (2017) are sufficient for the current article, we will use the updated fragments from Maestrini and Wand (2020) since they are the new standard for approximate Bayesian inference on variance and covariance matrix parameters.

3 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 function $\sigma(t,s) = \mathbb{E}[\{y_i(t)-\mu(t)\}\{y_i(s)-\mu(s)\}]$, $i = 1,\ldots,n$. Then, the covariance operator Σ of y_i is defined as

$$(\Sigma f)(t) \equiv \int_0^1 \sigma(t,s) f(s) ds, \quad f \in L^2[0,1]. \tag{14}$$

Mercer's Theorem implies that 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):

$$y_i(t) = \mu(t) + \sum_{l=1}^{\infty} \zeta_{il}^* \, \psi_l^*(t), \quad i = 1,\ldots,n, \tag{15}$$

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 $\mathbb{V}\text{ar}(\zeta_{il}^*) = \gamma_l$.

Expansion (15) 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 basis functions $\psi_1^*,\ldots,\psi_L^*$. That is, for any choice of L orthonormal basis functions $\psi_1,\ldots,\psi_L$, the minimum of

$$\sum_{i=1}^{n} ||y_i - \mu - \sum_{l=1}^{L} \langle y_i - \mu, \psi_l \rangle \psi_l||^2$$

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

$$\widehat{y}_i(t) \equiv \mu(t) + \sum_{l=1}^{L} \zeta_{il}^* \, \psi_l^*(t), \quad i = 1, \ldots, n. \tag{16}$$

Next, we make some observations involving the scores and the orthonormal basis functions in (15) and (16):

1. If $\gamma_l = \gamma_k$, $l \neq k$, then the corresponding eigenfunctions ψ_l^* and ψ_k^* are not unique. We will simply assume that the eigenvalues are unique. This is a reasonable assumption for most applications (e.g. climate data and biostatistical problems), where equal eigenvalues are normally encountered after the infinite series in (15) is truncated.

2. If all the eigenvalues are unique, the corresponding eigenfunctions are only unique up to a change of sign. 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 or the interpretation of the contribution of the eigenfunction. Here, we simply assume that the inner product of the Bayesian estimate of an eigenfunction and the eigenfunction itself is positive.

We state these assumptions formally.

Assumption 1. *The eigenvalues of the covariance operator in* (14) *are unique.*

Assumption 2. *The signs of the orthonormal basis functions* $\psi_1^*, \ldots, \psi_L^*$ *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 (16) are also possible, where

$$\widehat{y}_i(t) \equiv \mu(t) + \sum_{l=1}^{L} \zeta_{il} \, \psi_l(t), \quad i = 1, \ldots, n, \tag{17}$$

where ζ_{il} are correlated across l, but remain independent across i, and the ψ_l are not orthonormal. Theorem 3.1 shows that an orthogonal decomposition of the resulting basis functions and scores is sufficient for establishing the appropriate estimator (16). Its proof is provided in Appendix A.

Theorem 3.1. *Consider Assumptions 1 and 2 and the approximations of the response curves* $y_1, \ldots, y_n$ *in* (17)*. Then, there exists a unique set of orthogonal basis functions* $\widehat{\psi}_1, \ldots, \widehat{\psi}_L$ *and an uncorrelated set of scores* $\widehat{\zeta}_{i1}, \ldots, \widehat{\zeta}_{iL}$*,* $i = 1, \ldots, n$*, such that*

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

3.1 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 observations in time are not necessarily the same for each curve. Let the set of design points for the ith curve be summarized by the vector

$$\boldsymbol{t}_i \equiv (t_{i1}, \ldots, t_{iT_i})^\mathsf{T}, \quad i = 1, \ldots, n, \tag{18}$$

where T_i is the number of observations on the ith curve. In addition, we represent the observations for the ith curve, $y_i(t)$, by the vector

$$\boldsymbol{y}_i \equiv \{y_i(t_{i1}) + \varepsilon_{i1}, \ldots, y_i(t_{iT_i}) + \varepsilon_{iT_i}\}^\mathsf{T} \quad i = 1, \ldots, n, \tag{19}$$

where 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 (16) takes the form:

$$\boldsymbol{y}_i = \boldsymbol{\mu}_i + \sum_{l=1}^{L} \zeta_{il}\boldsymbol{\psi}_{il} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{20}$$

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

We model continuous curves from discrete observations via semiparametric 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 basis functions are:

$$\mu(t) \approx \beta_{\mu,0} + \beta_{\mu,1}t + \sum_{k=1}^{K} u_{\mu,k}z_k(t) \quad \text{and} \quad \psi_l(t) \approx \beta_{\psi_l,0} + \beta_{\psi_l,1}t + \sum_{k=1}^{K} u_{\psi_l,k}z_k(t) \quad \text{for } l = 1, \ldots, L,$$

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:

$$\boldsymbol{\beta}_\mu \equiv (\beta_{\mu,0}, \beta_{\mu,1})^\mathsf{T} \quad \boldsymbol{u}_\mu \equiv (u_{\mu,1}, \ldots, u_{\mu,K})^\mathsf{T} \quad \boldsymbol{\nu}_\mu \equiv (\boldsymbol{\beta}_\mu^\mathsf{T}, \boldsymbol{u}_\mu^\mathsf{T})^\mathsf{T}$$
$$\boldsymbol{\beta}_{\psi_l} \equiv (\beta_{\psi_l,0}, \beta_{\psi_1,1})^\mathsf{T}, \quad \boldsymbol{u}_{\psi_l} \equiv (u_{\psi_l,1}, \ldots, u_{\psi_l,K})^\mathsf{T} \quad \text{and} \quad \boldsymbol{\nu}_{\psi_l} \equiv (\boldsymbol{\beta}_{\psi_l}^\mathsf{T}, \boldsymbol{u}_{\psi_l}^\mathsf{T})^\mathsf{T} \quad \text{for } l = 1, \ldots, L.$$

Then simple derivations that stem from (20) show that the vector of observations on each of the response curves satisfies the representation:

$$\boldsymbol{y}_i = \boldsymbol{C}_i \left(\boldsymbol{\nu}_\mu + \sum_{l=1}^{L} \zeta_{il}\boldsymbol{\nu}_{\psi_l} \right) + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n,$$

where

$$\boldsymbol{C}_i \equiv \begin{bmatrix} 1 & t_{i1} & z_1(t_{i1}) & \dots & z_K(t_{i1}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{iT_i} & z_1(t_{iT_i}) & \dots & z_K(t_{iT_i}) \end{bmatrix}. \tag{21}$$

In addition, we define:

$$\boldsymbol{y} \equiv (\boldsymbol{y}_1^\intercal, \dots, \boldsymbol{y}_N^\intercal)^\intercal, \quad \boldsymbol{\nu} \equiv (\boldsymbol{\nu}_\mu^\intercal, \boldsymbol{\nu}_{\psi_1}^\intercal, \dots, \boldsymbol{\nu}_{\psi_L}^\intercal)^\intercal \quad \text{and} \quad \boldsymbol{\zeta}_i \equiv (\zeta_{i1}, \dots, \zeta_{iL})^\intercal \quad i = 1, \dots, n. \tag{22}$$

Next, we present the Bayesian FPCA Gaussian response model:

$$\begin{gathered} \boldsymbol{y}_i | \boldsymbol{\nu}, \boldsymbol{\zeta}_i, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} \mathrm{N}\left\{ \boldsymbol{C}_i \left(\boldsymbol{\nu}_\mu + \sum_{l=1}^{L} \zeta_{il} \boldsymbol{\nu}_{\psi_l} \right), \sigma_\varepsilon^2 \boldsymbol{I}_{T_i} \right\}, \quad \boldsymbol{\zeta}_i \overset{\text{ind.}}{\sim} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\zeta_i}), \quad i = 1, \dots, n, \\ \begin{bmatrix} \boldsymbol{\nu}_\mu \\ \boldsymbol{\nu}_{\psi_l} \end{bmatrix} \Bigg| \, \sigma_\mu^2, \sigma_{\psi_l}^2 \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_\mu \\ \boldsymbol{\mu}_{\psi_l} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_\mu & \boldsymbol{O}^\intercal \\ \boldsymbol{O} & \boldsymbol{\Sigma}_{\psi_l} \end{bmatrix} \right), \quad \sigma_{\psi_l}^2 | a_{\psi_l} \overset{\text{ind.}}{\sim} \text{Inverse} - \chi^2(1, 1/a_{\psi_l}), \\ a_{\psi_l} \overset{\text{ind.}}{\sim} \text{Inverse} - \chi^2(1, 1/A_{\psi_l}^2), \quad l = 1, \dots, L, \\ \sigma_\mu^2 | a_\mu \sim \text{Inverse} - \chi^2(1, 1/a_\mu), \quad a_\mu \sim \text{Inverse} - \chi^2(1, 1/A_\mu^2), \\ \sigma_\varepsilon^2 | a_\varepsilon \sim \text{Inverse} - \chi^2(1, 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse} - \chi^2(1, 1/A_\varepsilon^2), \end{gathered} \tag{23}$$

where

$$\begin{gathered} \boldsymbol{\mu}_\mu \equiv (\boldsymbol{\mu}_{\beta_\mu}^\intercal, \boldsymbol{0}_K^\intercal)^\intercal, \quad \boldsymbol{\Sigma}_\mu \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\beta_\mu} & \boldsymbol{O}^\intercal \\ \boldsymbol{O} & \sigma_\mu^2 \boldsymbol{I}_K \end{bmatrix}, \\ \boldsymbol{\mu}_{\psi_l} \equiv (\boldsymbol{\mu}_{\beta_{\psi_l}}^\intercal, \boldsymbol{0}_K^\intercal)^\intercal, \quad \boldsymbol{\Sigma}_{\psi_l} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\beta_{\psi_l}} & \boldsymbol{O}^\intercal \\ \boldsymbol{O} & \sigma_{\psi_l}^2 \boldsymbol{I}_K \end{bmatrix}, \quad l = 1, \dots, L, \end{gathered} \tag{24}$$

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

Full Bayesian inference for the parameter set $\boldsymbol{\nu}$, $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_n$, σ_ε^2, a_ε, σ_μ^2, a_μ, $\sigma_{\psi_1}^2, \dots, \sigma_{\psi_L}^2$ and $a_{\psi_1}, \dots, a_{\psi_L}$ requires the determination of the posterior density function

$$p(\boldsymbol{\nu}, \boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_n, \sigma_\varepsilon^2, a_\varepsilon, \sigma_\mu^2, a_\mu, \sigma_{\psi_1}^2, \dots, \sigma_{\psi_L}^2, a_{\psi_1}, \dots, a_{\psi_L} | \boldsymbol{y}),$$

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 (23), even for moderate dimensions for $\boldsymbol{\nu}$, which

depends on the number of FPC basis functions (L) and O'Sullivan penalised spline basis functions (K). Second, the mean function $\mu(t)$ and the FPC basis functions $\psi_1(t),\ldots,\psi_L(t)$ are typically highly correlated, which is expected to lead to poor mixing. After preliminary MCMC simulations through Rstan, the R (R Core Team, 2020) interface to the probabilistic programming language Stan (Stan Development Team, 2020), we found that poor mixing resulted in posterior variance and covariance estimates that were unreliable.

Alternatively, variational approximate inference for model (23) involves the mean field restriction:

$$p(\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon,a_\varepsilon,\sigma^2_\mu,a_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L},a_{\psi_1},\ldots,a_{\psi_L}|\boldsymbol{y}) \approx \left\{\prod_{i=1}^N q(\boldsymbol{\zeta}_i)\right\} q(\boldsymbol{\nu},a_\varepsilon,a_\mu,a_{\psi_1},\ldots,a_{\psi_L}) q(\sigma^2_\varepsilon,\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L}). \tag{25}$$

The approximation in (25) represents the minimal mean-field restriction that is required for approximate variational inference. Here, we have assumed posterior independence between global parameters and response curve-specific parameters, 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 (25) becomes

$$\left\{\prod_{i=1}^N q(\boldsymbol{\zeta}_i)\right\} q(\boldsymbol{\nu}) q(\sigma^2_\varepsilon) q(a_\varepsilon) q(\sigma^2_\mu) q(a_\mu) \left\{\prod_{l=1}^L q(\sigma^2_{\psi_l}) q(a_{\psi_l})\right\}. \tag{26}$$

From here, we work with the factorization in (26) to minimize the Kullback-Leibler divergence of the right-hand side of (25) from its left-hand side. The factor graph for model (23) that represents the factorization in (26) is presented in Figure 2.

From Figure 2, we identify the following factor graph fragments that are involved in the Bayesian FPCA model (23):

- The fragments for $p(\boldsymbol{\zeta}_1),\ldots,p(\boldsymbol{\zeta}_n)$ are Gaussian prior fragments. The updates for this fragment are presented in Section 4.1.1 of Wand (2017), where a vec-based representation of the Multivariate Normal density function is used.

- The fragments for $p(a_\varepsilon)$, $p(a_\mu)$ and $p(a_{\psi_1}),\ldots,p(a_{\psi_L})$ are scalar versions of the Inverse G-Wishart prior fragment, which is presented as Algorithm 1 of Maestrini and Wand (2020).

- The fragments for $p(\sigma^2_\varepsilon|a_\varepsilon)$, $p(\sigma^2_\mu|a_\mu)$ and $p(\sigma^2_{\psi_1}|a_{\psi_1}),\ldots,p(\sigma^2_{\psi_L}|a_{\psi_L})$ are scalar versions of the iterated Inverse G-Wishart fragment, which is presented as Algorithm 2 of Maestrini and Wand (2020).

- The fragments for $p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)$ and $p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})$ are two new fragments that have not been addressed in previous literature on VMP, but are crucial for FPCA modelling. We name the fragment for $p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)$ the *FPCA Gaussian likelihood fragment* and the fragment for $p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})$ the *mean and FPC Gaussian penalization fragment*.

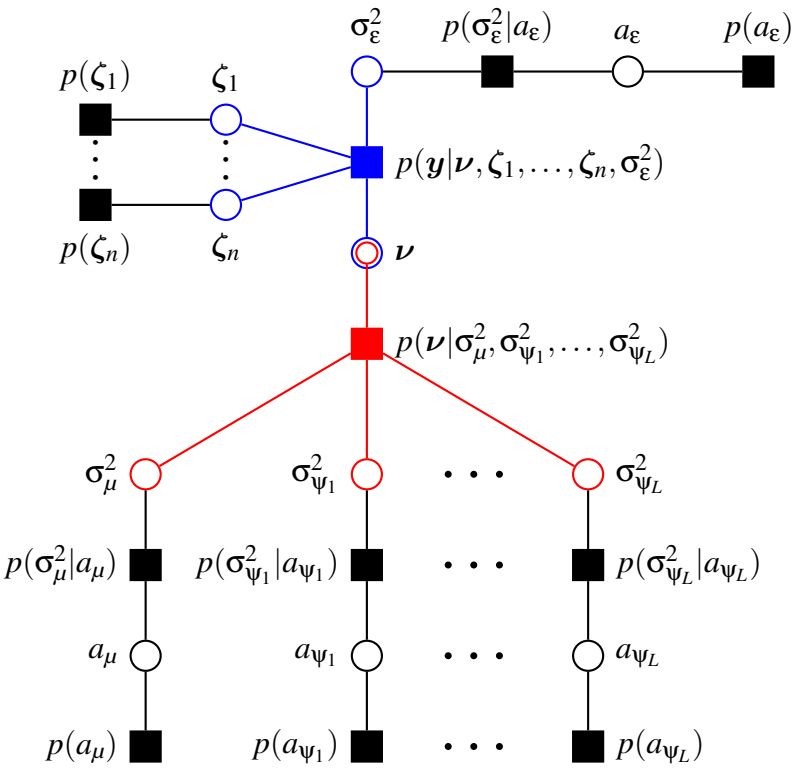


Figure 2: The factor graph for the Bayesian FPCA model in (23).

3.2 FPCA Gaussian Likelihood Fragment

The FPCA Gaussian likelihood fragment, shown in blue in Figure 2, is defined by the factor

$$p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)=\prod_{i=1}^{n}p(\boldsymbol{y}_i|\boldsymbol{\nu},\zeta_i,\sigma^2_\varepsilon), \tag{27}$$

where

$$\boldsymbol{y}_i|\boldsymbol{\nu},\zeta_i,\sigma^2_\varepsilon \overset{\text{ind.}}{\sim} \mathrm{N}\left\{\boldsymbol{C}_i\left(\boldsymbol{\nu}_\mu+\sum_{l=1}^{L}\zeta_{il}\boldsymbol{\nu}_{\psi_l}\right),\sigma^2_\varepsilon\boldsymbol{I}_{T_i}\right\},\quad \text{for } i=1,\ldots,n. \tag{28}$$

The purpose of this fragment is to provide message updates for the variational posterior density functions of $\boldsymbol{\nu}$, $\zeta_i,\ldots,\zeta_n$ and σ^2_ε at every iteration of the VMP algorithm. Here, we outline the construction of the messages that are passed from the factor representing the likelihood $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)$ to each of its neighbouring stochastic nodes. On the other hand, the derivations of these messages and the derivations of expected values of random variables, random vectors and random matrices that these messages depend on are deferred to Appendix D.

The message from $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)$ to $\boldsymbol{\nu}$ can be shown to be proportional to a Multivariate Normal density function, with vec-based exponential density function representation:

$$m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}}(\boldsymbol{\nu})\propto\exp\left\{\begin{bmatrix}\boldsymbol{\nu}\\ \mathrm{vec}(\boldsymbol{\nu}\boldsymbol{\nu}^\intercal)\end{bmatrix}^\intercal\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}}\right\}. \tag{29}$$

The update for the natural parameter vector in (29) is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}} \longleftarrow \begin{bmatrix} \mathbb{E}_q(1/\sigma^2_\varepsilon)\sum_{i=1}^{n}\left\{\mathbb{E}_q(\widetilde{\zeta}_i)^\intercal \otimes \boldsymbol{C}_i\right\}^\intercal \boldsymbol{y}_i \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2_\varepsilon)\sum_{i=1}^{n}\operatorname{vec}\left\{\mathbb{E}_q(\widetilde{\zeta}_i\widetilde{\zeta}_i^\intercal)\otimes(\boldsymbol{C}_i^\intercal\boldsymbol{C}_i)\right\} \end{bmatrix}, \tag{30}$$

where,

$$\widetilde{\zeta}_i \equiv (1,\zeta_i^\intercal)^\intercal, \quad \text{for } i=1,\ldots,n. \tag{31}$$

Before proceeding to the other messages in this fragment, we make a brief comment on using the vec-based representation of the message in (29), as opposed to the storage-economical vech-based representation. 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 penalised spline basis functions ($K=25$) for semiparametric 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 of the message in (29), which does not require the use of a Moore-Penrose inverse matrix.

For each $i=1,\ldots,n$, the message from $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)$ to ζ_i is

$$m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\zeta_i}(\zeta_i) \propto \exp\left\{\begin{bmatrix}\zeta_i \\ \operatorname{vech}(\zeta_i\zeta_i^\intercal)\end{bmatrix}^\intercal \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\zeta_i}\right\}, \tag{32}$$

which is proportional to a Multivariate Normal density function. The update for the natural parameter vector in (32) is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\zeta_i} \longleftarrow \begin{bmatrix}\mathbb{E}_q(1/\sigma^2_\varepsilon)\left\{\mathbb{E}_q(\boldsymbol{V}_\psi)^\intercal\boldsymbol{C}_i^\intercal\boldsymbol{y}_i - \mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i})\right\} \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2_\varepsilon)\boldsymbol{D}_L^\intercal\operatorname{vec}\{\mathbb{E}_q(\boldsymbol{H}_{\psi,i})\}\end{bmatrix}, \tag{33}$$

where

$$\boldsymbol{V}_\psi \equiv \begin{bmatrix}\boldsymbol{\nu}_{\psi_1} & \ldots & \boldsymbol{\nu}_{\psi_L}\end{bmatrix} \quad \text{and} \quad \boldsymbol{h}_{\mu\psi,i} \equiv \boldsymbol{V}_\psi^\intercal\boldsymbol{C}_i^\intercal\boldsymbol{C}_i\boldsymbol{\nu}_\mu, \quad \boldsymbol{H}_{\psi,i} \equiv \boldsymbol{V}_\psi^\intercal\boldsymbol{C}_i^\intercal\boldsymbol{C}_i\boldsymbol{V}_\psi, \quad \text{for } i=1,\ldots,n. \tag{34}$$

The message from $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)$ to σ^2_ε is

$$m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon}(\sigma^2_\varepsilon) \propto \exp\left\{\begin{bmatrix}\log(\sigma^2_\varepsilon) \\ 1/\sigma^2_\varepsilon\end{bmatrix}^\intercal \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon}\right\}, \tag{35}$$

and it is proportional to an Inverse $-\chi^2$ density function. The update for the natural parameter vector in (35) is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon} \longleftarrow \begin{bmatrix} -\frac{1}{2}\sum_{i=1}^{n} T_i \\ -\frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_q\left\{\left(\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{V}\widetilde{\boldsymbol{\zeta}}_i\right)^\intercal\left(\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{V}\widetilde{\boldsymbol{\zeta}}_i\right)\right\} \end{bmatrix}, \tag{36}$$

where

$$\boldsymbol{V} \equiv \begin{bmatrix} \boldsymbol{\nu}_\mu & \boldsymbol{\nu}_{\psi_1} & \ldots & \boldsymbol{\nu}_{\psi_L} \end{bmatrix}. \tag{37}$$

We must remember that the Inverse$-\chi^2$ density function message that is passed to σ^2_ε is part of the Inverse G-Wishart class of density functions for VMP. Within this class of messages, a graph message is also required to specify whether the density function respects a full or a diagonal matrix. This graphical message does not affect Inverse$-\chi^2$ density function messages, however we will include a graph message with the aim of providing fragments that are compatible with previously constructed Inverse G-Wishart fragments (Maestrini & Wand, 2020, Algorithms 1 & 2). According to Section 7.4 of Maestrini and Wand (2020), the auxiliary-based hierarchical prior specification of σ^2_ε in (23) requires a graphical message of the form

$$G_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon} \longleftarrow G_{\text{full}}. \tag{38}$$

That is, the univariate random variable σ^2_ε is chosen to respect a "full" 1×1 matrix.

Pseudocode for the FPCA 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 FPCAGAUSSIANLIKELIHOODFRAGMENT

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 all expectations with respect to the optimal posterior distribution. ▷ see Appendix D
Update $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}}$ ▷ equation (30)
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 (33)
Update $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon}$ ▷ equation (36)
Update $G_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)\to\sigma^2_\varepsilon}$ ▷ equation (38)

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 Mean and FPC Gaussian Penalization Fragment

The mean and FPC Gaussian penalization fragment, shown in red in Figure 2, is defined by the factor

$$p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L}).$$

where

$$\begin{bmatrix} \boldsymbol{\nu}_{\mu} \\ \boldsymbol{\nu}_{\psi_l} \end{bmatrix} \Bigg| \; \sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L} \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\mu} \\ \boldsymbol{\mu}_{\psi_l} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\mu} & \mathbf{O}^{\mathsf{T}} \\ \mathbf{O} & \boldsymbol{\Sigma}_{\psi_l} \end{bmatrix} \right), \quad \text{for } l = 1, \ldots, L. \tag{39}$$

and all sub-vectors and sub-matrices are defined in (24). The purpose of this fragment is to provide message updates for the variational posterior density functions of $\boldsymbol{\nu}$, σ^2_{μ} and $\sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}$ at each iteration of the VMP algorithm. Here, as in Section 3.2, we outline the messages that are passed from this factor to its neighbouring stochastic nodes. For detailed derivations of these messages and all relevant expectations, we defer the reader to Appendix E.

First, let us introduce the vector and matrix

$$\boldsymbol{\mu}_{\mathrm{v}} \equiv (\boldsymbol{\mu}_{\mu}^{\mathsf{T}}, \boldsymbol{\mu}_{\psi_1}^{\mathsf{T}}, \ldots, \boldsymbol{\mu}_{\psi_L}^{\mathsf{T}})^{\mathsf{T}} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathrm{v}} \equiv \mathrm{blockdiag}(\boldsymbol{\Sigma}_{\mu}, \boldsymbol{\Sigma}_{\psi_1}, \ldots, \boldsymbol{\Sigma}_{\psi_L}). \tag{40}$$

Then, the message from $p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L})$ to $\boldsymbol{\nu}$ can be shown to be proportional to a Multivariate Normal density function, with vec-based exponential density representation

$$m_{p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) \to \boldsymbol{\nu}}(\boldsymbol{\nu}) \propto \exp\left\{ \begin{bmatrix} \boldsymbol{\nu} \\ \mathrm{vec}(\boldsymbol{\nu}\boldsymbol{\nu}^{\mathsf{T}}) \end{bmatrix}^{\mathsf{T}} \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) \to \boldsymbol{\nu}} \right\}, \tag{41}$$

where

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) \to \boldsymbol{\nu}} \longleftarrow \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\Sigma}_{\mathrm{v}}^{-1}) \boldsymbol{\mu}_{\mathrm{v}} \\ -\frac{1}{2} \mathrm{vec}\left\{ \mathbb{E}_q(\boldsymbol{\Sigma}_{\mathrm{v}}^{-1}) \right\} \end{bmatrix}. \tag{42}$$

Once again, we have used a vec-based representation of the message to $\boldsymbol{\nu}$ as opposed to a storage-economical vech-based representation. The major reason for this is outlined in the discussion following (31).

The message from $p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L})$ to σ^2_{μ} is

$$m_{p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) \to \sigma^2_{\mu}}(\sigma^2_{\mu}) \propto \exp\left\{ \begin{bmatrix} \log(\sigma^2_{\mu}) \\ 1/\sigma^2_{\mu} \end{bmatrix}^{\mathsf{T}} \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) \to \sigma^2_{\mu}} \right\}, \tag{43}$$

which is an Inverse $-\chi^2$ density function after normalization. The update for the natural parameter vector in (43) is

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) \to \sigma^2_{\mu}} \longleftarrow \begin{bmatrix} -\frac{K}{2} \\ -\frac{1}{2} \mathbb{E}_q(\boldsymbol{u}_{\mu}^{\mathsf{T}} \boldsymbol{u}_{\mu}) \end{bmatrix}. \tag{44}$$

For $l = 1, \ldots, L$, the message from $p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L})$ to $\sigma^2_{\psi_l}$ is similar to the message to σ^2_{μ}. The message is

$$m_{p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) \to \sigma^2_{\mu}}(\sigma^2_{\psi_l}) \propto \exp\left\{ \begin{bmatrix} \log(\sigma^2_{\psi_l}) \\ 1/\sigma^2_{\psi_l} \end{bmatrix}^{\mathsf{T}} \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_{\mu}, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) \to \sigma^2_{\psi_l}} \right\}, \tag{45}$$

which is an Inverse $-\chi^2$ density function after normalization. The update for the natural parameter vector in (45) is

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\psi_l}} \longleftarrow \begin{bmatrix} -\frac{K}{2} \\ -\frac{1}{2}\,\mathbb{E}_q(\boldsymbol{u}_{\psi_l}^{\mathsf{T}}\boldsymbol{u}_{\psi_l}) \end{bmatrix}. \tag{46}$$

Finally, recall the discussion following (37). Each of the messages to the variance parameters $\sigma^2_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}$ must be paired with a graph message. For the same reasons that were used to justify the graphical message in (38), the graph messages received by $\sigma^2_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}$ are, respectively,

$$G_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_\mu} \longleftarrow G_{\text{full}} \quad \text{and} \quad G_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\psi_l}} \longleftarrow G_{\text{full}}, \quad \text{for } l = 1,\ldots,L. \tag{47}$$

Pseudocode for the mean and FPC 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 MEANANDFPCGAUSSIANPENALIZATIONFRAGMENT

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

Updates:

Update all expectations with respect to the optimal posterior distribution. ▷ 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 (42)
Update $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_\mu}$ ▷ equation (44)
Update $G_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_\mu}$ ▷ equation (47)
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 (46)
    Update $G_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\sigma^2_{\psi_l}}$ ▷ equation (47)

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

4 Post-VMP Steps

The FPCA model for curve estimation (16), which has its genesis in the Karhunen-Loève decomposition (15), relies on orthogonal functional principal component basis functions 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 basis functions. Although curve estimation is still valid without these constraints, interpretation of the analysis is more straightforward with orthogonal basis functions. Furthermore, the basis functions 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 basis functions $\boldsymbol{\nu}$ and the vectors of principal component scores $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$. Upon convergence of the algorithm, the natural parameter vectors for these optimal posterior density functions are, according to (9),

$$\boldsymbol{\eta}_{q^*(\boldsymbol{\nu})} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)\to\boldsymbol{\nu}} + \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})\to\boldsymbol{\nu}}$$

and

$$\boldsymbol{\eta}_{q^*(\boldsymbol{\zeta}_i)} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)\to\boldsymbol{\zeta}_i} + \boldsymbol{\eta}_{p(\boldsymbol{\zeta}_i)\to\boldsymbol{\zeta}_i}, \quad \text{for } i = 1, \ldots, n.$$

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 (12), 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 (13). Note that we partition $\mathbb{E}_q(\boldsymbol{\nu})$ as

$$\mathbb{E}_q(\boldsymbol{\nu}) = \{\mathbb{E}_q(\boldsymbol{\nu}_\mu)^\intercal, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_1})^\intercal, \ldots, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_L})^\intercal\}^\intercal$$

in a similar fashion to (22).

4.2 Establishing a Vector Version of the Karhunen-Loève Decomposition

In this section, we outline a sequence of steps to establish orthogonal functional principal component basis functions and uncorrelated scores. Note that we will treat the estimated functional principal component basis functions 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 functional principal component scores (e.g. Yao et al., 2005; Benko, Härdle, & Kneip, 2009).

Now, we outline the steps to construct orthogonal functional principal component basis functions and uncorrelated scores. The existence and uniqueness of the basis functions are justified by Theorem 3.1. 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 (21):

$$\boldsymbol{C}_g \equiv \begin{bmatrix} 1 & t_{g1} & z_1(t_{g1}) & \ldots & z_K(t_{g1}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{gn_g} & z_1(t_{gn_g}) & \ldots & z_K(t_{gn_g}) \end{bmatrix}.$$

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 basis functions $\mathbb{E}_q\{\psi_l(\boldsymbol{t}_g)\} = \boldsymbol{C}_g\,\mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})$, $l = 1, \ldots, L$. Then define the matrix $\boldsymbol{\Psi}$ such that

$$\boldsymbol{\Psi} \equiv \begin{bmatrix} \mathbb{E}_q\{\psi_1(t_g)\} & \cdots & \mathbb{E}_q\{\psi_L(t_g)\} \end{bmatrix}.$$

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

$$\boldsymbol{\Xi} \equiv \begin{bmatrix} \mathbb{E}_q(\zeta_1) & \cdots & \mathbb{E}_q(\zeta_n) \end{bmatrix}^\intercal$$

Set $\boldsymbol{m}_\zeta$ to be the $L \times 1$ sample mean vector of the columns of $\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal \boldsymbol{\Xi}^\intercal$, and set

$$\widehat{\mu}(\boldsymbol{t}_g) \equiv \mathbb{E}_q\{\mu(\boldsymbol{t}_g)\} + \boldsymbol{U}_\psi \boldsymbol{m}_\zeta. \tag{48}$$

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

$$\widetilde{\boldsymbol{\Psi}} \equiv \boldsymbol{U}_\psi \boldsymbol{Q} \boldsymbol{\Lambda}^{1/2} \quad \text{and} \quad \widetilde{\boldsymbol{\Xi}} \equiv (\boldsymbol{\Xi} \boldsymbol{V}_\psi \boldsymbol{D}_\psi - \boldsymbol{1}_n \boldsymbol{m}_\zeta^\intercal) \boldsymbol{Q} \boldsymbol{\Lambda}^{-1/2}. \tag{49}$$

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

$$\widetilde{\boldsymbol{\Psi}} = \begin{bmatrix} \widetilde{\psi}_1(\boldsymbol{t}_g) & \cdots & \widetilde{\psi}_L(\boldsymbol{t}_g) \end{bmatrix} \quad \text{and} \quad \widetilde{\boldsymbol{\Xi}} = \begin{bmatrix} \widetilde{\zeta}_{11} & \cdots & \widetilde{\zeta}_{1L} \\ \vdots & \ddots & \vdots \\ \widetilde{\zeta}_{N1} & \cdots & \widetilde{\zeta}_{NL} \end{bmatrix}$$

The columns of $\widetilde{\boldsymbol{\Psi}}$ are orthogonal 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 (16) over the vector $\boldsymbol{t}_g$ with

$$\widehat{\psi}_l(\boldsymbol{t}_g) \equiv \frac{\widetilde{\psi}_l(\boldsymbol{t}_g)}{||\widetilde{\psi}_l||}, \quad l = 1, \ldots, L, \tag{50}$$

as well as estimates of the scores with

$$\widehat{\zeta}_{il} \equiv ||\widetilde{\psi}_l|| \widetilde{\zeta}_{il}, \quad i = 1, \ldots, n, \quad l = 1, \ldots, L.$$

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*

$$\widehat{y}_i(\boldsymbol{t}_g) = \widehat{\mu}(\boldsymbol{t}_g) + \sum_{l=1}^{L} \widehat{\zeta}_{il}\widehat{\psi}_l(\boldsymbol{t}_g), \quad i = 1,\ldots,n. \tag{51}$$

Proposition 4.2. *The vectors $\widehat{\boldsymbol{\zeta}}_1,\ldots,\widehat{\boldsymbol{\zeta}}_N$ are independent and satisfy:*

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\zeta}}_i = \mathbf{0} \quad \textit{and} \quad \frac{1}{n-1}\sum_{i=1}^{n}\widehat{\boldsymbol{\zeta}}_i\widehat{\boldsymbol{\zeta}}_i^{\intercal} = \operatorname{diag}\left(||\widetilde{\psi}_1||^2,\ldots,||\widetilde{\psi}_L||^2\right).$$

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 (15). Furthermore, the vectors $\psi_1^*(\boldsymbol{t}_g),\ldots,\psi_L^*(\boldsymbol{t}_g)$ respect the orthogonality conditions in ℓ^2. Therefore, (51) 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 3).

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 (23). Note that the VMP iterative loop has the following generic steps (Minka, 2005; Wand, 2017):

1. Choose a factor.
2. Update the parameter vectors of the messages passed from the factor's neighbouring stochastic nodes to the factor.
3. Update the parameter vectors of the messages passed from the factor to its neighbouring stochastic nodes.

These steps are summarized in Algorithm 3.

5.1 Accuracy Assessment

For model (23), we simulated 36 response curves with the number of observations T_i for the ith curve sampled uniformly over $\{20,21,\ldots,30\}$. Furthermore, the observations within the ith curve $\{t_{i1},\ldots,t_{iT_i}\}$ were sampled uniformly over the interval $(0,1)$, while the residual variance σ_ε^2 was set to 1. The mean function was

$$\mu(t) = 3\sin(\pi t), \tag{52}$$

Algorithm 3 Generic VMP algorithm for Gaussian response FPCA models.

Inputs: All hyperparameters and observed data
Initialize: All factor to stochastic node messages. ▷ see (6)
Updates:

while $\log \underline{p}(\boldsymbol{y};q)$ has not converged **do**
  Update all stochastic node to factor messages. ▷ see (7)
  Update the fragment for $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma^2_\varepsilon)$ ▷ see Algorithm 1
  Update the fragment for $p(\sigma^2_\varepsilon|a_\varepsilon)$ ▷ see Algorithm 2 of Maestrini and Wand (2020)
  Update the fragment for $p(a_\varepsilon)$ ▷ see Algorithm 1 of Maestrini and Wand (2020)
  Update the fragment for $p(\boldsymbol{\nu}|\sigma^2_\mu,\sigma^2_{\psi_1},\ldots,\sigma^2_{\psi_L})$ ▷ see Algorithm 2
  for $i = 1,\ldots,n$ **do**
    Update the fragment for $p(\zeta_i)$ ▷ see Section 4.1.1 of Wand (2017)
  Update the fragment for $p(\sigma^2_\mu|a_\mu)$ ▷ see Algorithm 2 of Maestrini and Wand (2020)
  Update the fragment for $p(a_\mu)$ ▷ see Algorithm 1 of Maestrini and Wand (2020)
  for $i = 1,\ldots,n$ **do**
    Update the fragment for $p(\sigma^2_{\psi_l}|a_{\psi_l})$ ▷ see Algorithm 2 of Maestrini and Wand (2020)
    Update the fragment for $p(a_{\psi_l})$ ▷ see Algorithm 1 of Maestrini and Wand (2020)
  Rotate, translate and re-scale $\boldsymbol{\Psi}$ and $\boldsymbol{\Xi}$. ▷ see Section 4.2

Outputs: $\widehat{\mu}(\boldsymbol{t}_g)$, $\widehat{\psi}_1(\boldsymbol{t}_g),\ldots,\widehat{\psi}_L(\boldsymbol{t}_g)$ and $\widehat{\zeta}_1,\ldots,\widehat{\zeta}_n$.

and the functional principal component basis functions were

$$\psi_1(t) = \sqrt{2}\sin(2\pi t) \quad \text{and} \quad \psi_2(t) = \sqrt{2}\cos(2\pi t), \tag{53}$$

where $L = 2$. Each vector of principal component scores were simulated according to

$$\zeta_i = \begin{bmatrix} \zeta_{i1} \\ \zeta_{i2} \end{bmatrix} \overset{\text{ind.}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix} \right), \quad i = 1,\ldots,n. \tag{54}$$

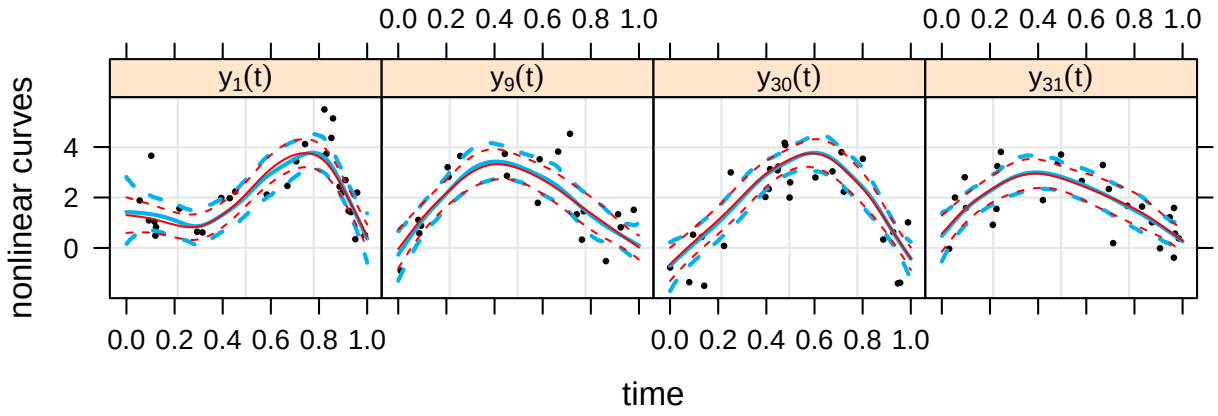
Semiparameteric regression with O'Sullivan penalized splines for the nonlinear curves was performed with $K = 10$.

The results from the simulation are presented in Figure 3, 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 (23) in blue for comparison with the VMP-based variational Bayes fits in red. 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.

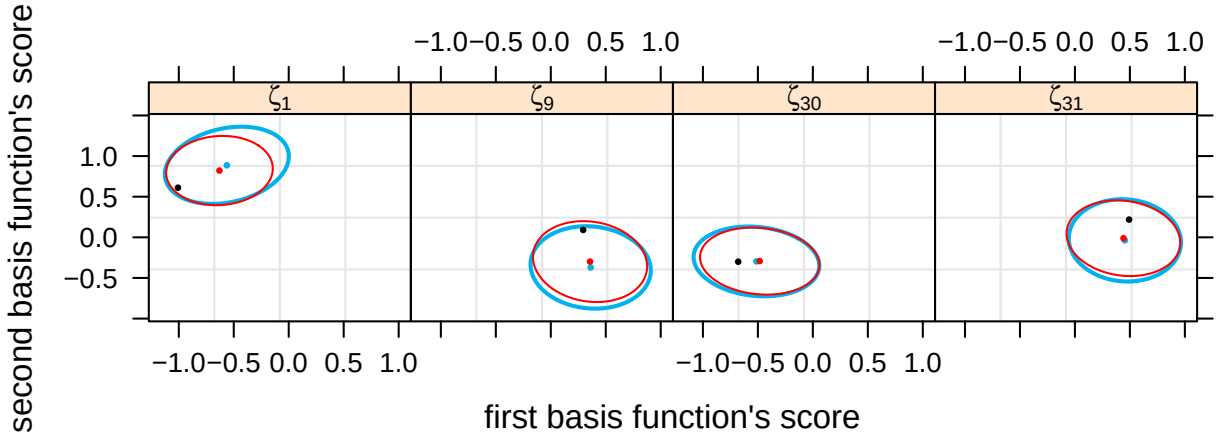
We then incorporated three settings for the number of response curves: $N \in \{10, 50, 100\}$. For each of these settings, we conducted 100 simulations of model (23) with the aim of analysing the error of the posterior mean estimates of the mean curve in (52) and the functional principal component basis functions in (53). The error of each simulation was determined via the integrated squared error:

$$\text{ISE}(f,\widehat{f}) = \int_0^1 \left| f(x) - \widehat{f}(x) \right|^2 dx, \tag{55}$$

where, in our simulations, $f(\cdot)$ represents the true function that generated the data, while $\widehat{f}(\cdot)$ repre-



(a)

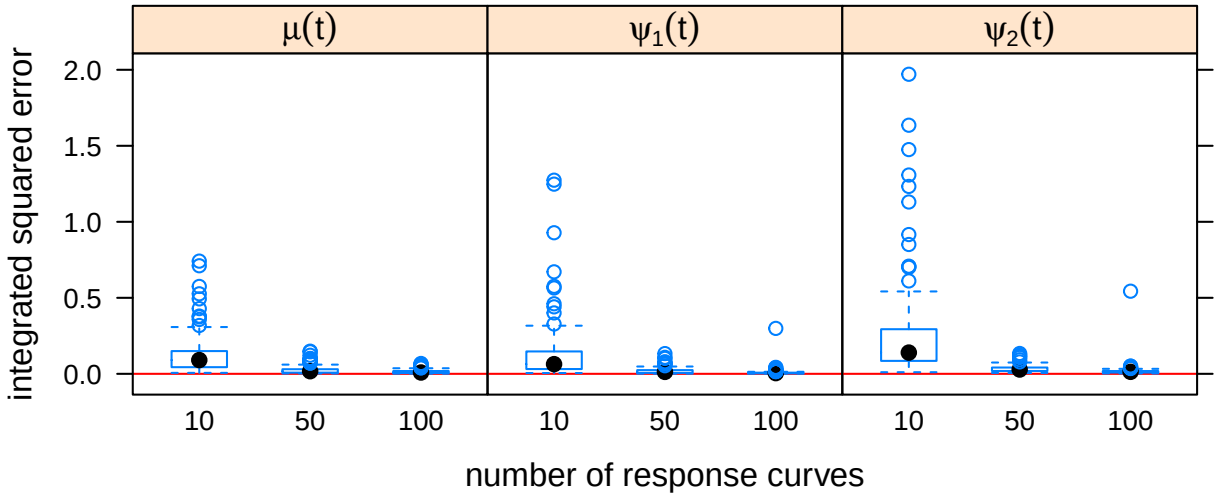


(b)

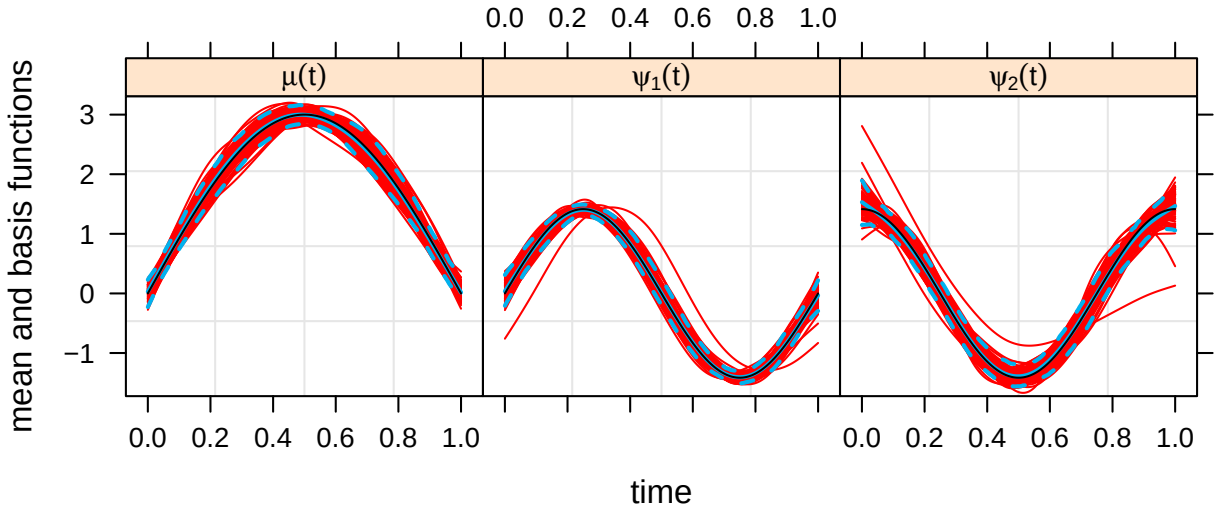
Figure 3: The results from one simulation of the Gaussian response FPCA model in (23). The simulation parameters are outlined in Section 5. In (a), the simulated data is shown in black, while the VMP-based variational Bayes posterior estimates are presented in red and the corresponding MCMC estimates 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.

sents the VMP-based variational Bayes posterior mean curve.

The error scores in Figure 4 (a) reflect the excellent results for the settings where $N = 50$ and 100. 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 basis function $\psi_2(\cdot)$ are poor. This is to be expected because the functional principal component scores associated with this basis function were generated from a $\mathrm{N}(0, 0.25)$ distribution reflecting its weaker contribution to the data generation process. Also, as expected, the error scores for all curves tend to decline with increasing N. In Figure 4 (b), we present all of the simulated posterior mean curves for the case where $N = 100$ for the mean function and the functional principal component basis functions, which are overlaid with



(a)



(b)

Figure 4: The results from a simulation study of the Gaussian response FPCA model in (23). The simulation parameters are outlined in Section 5.1. The box plots in (a) are a summary of the error scores, determined by the integrated squared error (55), for 100 simulations of each of the settings $N \in \{10, 50, 100\}$. In (b), we present the mean function (52) and the functional principal component basis functions (53) in black. For the case where $N = 100$, we show the resulting VMP-based posterior mean curve for each of the simulations in red. In addition, we have included the pointwise mean curve and the pointwise 95% confidence intervals resulting from the MCMC posterior estimates for each of the generated datasets for $N = 100$ in blue.

the true functions in black. These plots demonstrate the strength of the VMP algorithm for estimating the underlying curves that generate an observed set of functional data. 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 for each of the generated data sets. Evidently, there is strong agreement between the VMP simulations and the MCMC simulations in this setting.

Table 1: Median (first quartile, third quartile) elapsed computing time in seconds for VMP and MCMC with $N \in (10, 50, 100)$. 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	3.7 (2.3, 4.8)	83.7 (75.1, 93.1)	22.6
50	10.6 (6.5, 18.7)	419.2 (381.3, 453.8)	39.5
100	23.6 (13.1, 34.9)	1107.7 (1009.3, 1179.4)	46.9

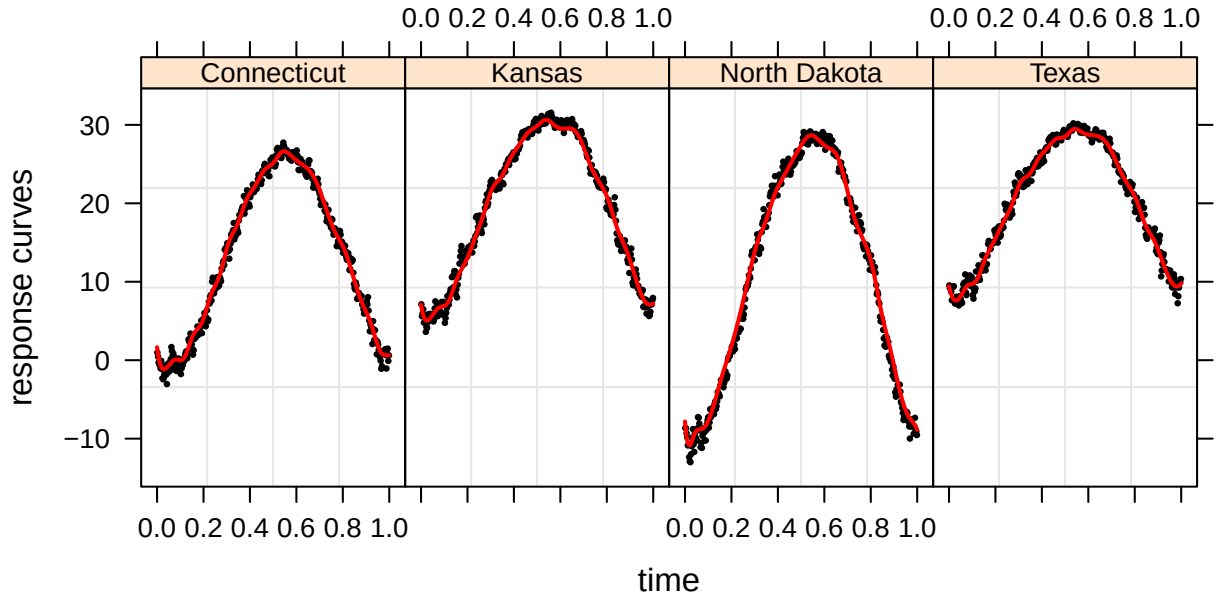
5.2 Computational Speed Comparisons

In the previous section, we saw that the mean field product restriction in (26) 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, Lee and Wand (2016) and Nolan et al. (2020) established speed gains for multilevel data models with streamlined matrix algebraic results and Luts and Wand (2015) presented timing comparisons for semiparametric regression models with count responses. In all cases, the variational Bayesian inference algorithms had strong accuracy in comparison to MCMC simulations and were far superior in computational speed.

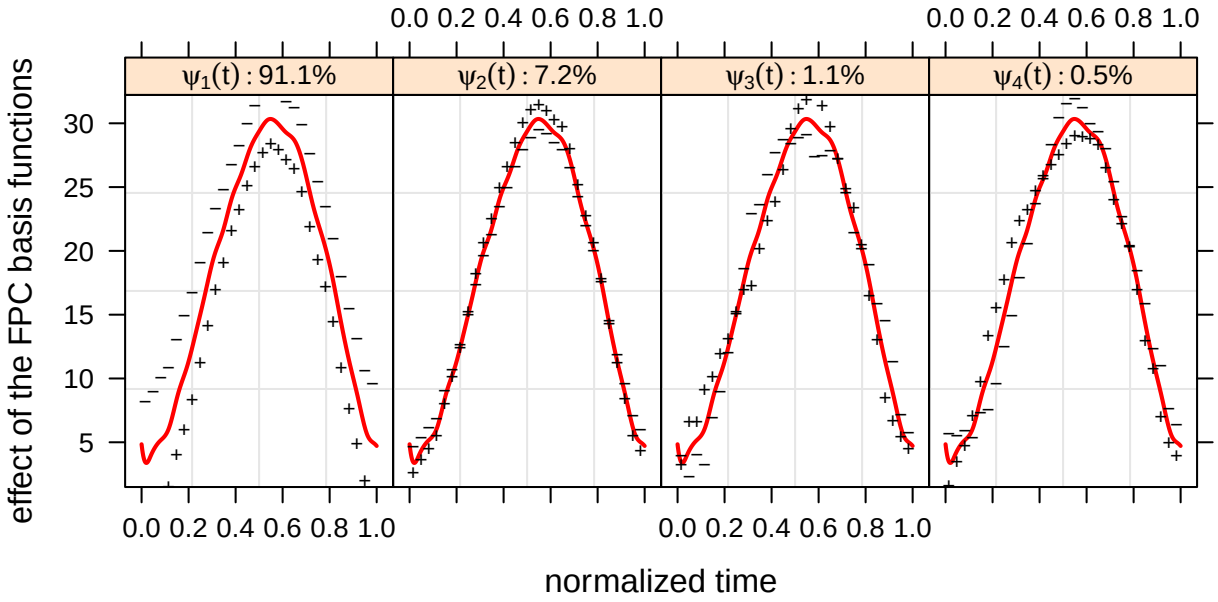
In Table 1, we present a similar set of results for the computational speed of VMP and MCMC for model (23). The simulations were identical to those that were used to generate the results in Figure 4, where there were 100 simulations over three settings for the number of response curves $N \in \{10, 50, 100\}$. 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 20 minutes, for $N = 100$. The most impressive results are in the fourth column, where the median VMP simulation is 22.6 times faster than the median MCMC simulation for $N = 10$, 39.5 times faster for $N = 50$ and 46.9 times faster for $N = 100$.

6 Application: United States Weather Data

We now provide an illustration of our methodology with an application to weather 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 interace 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.



(a)



(b)

Figure 5: Application of the VMP algorithm for FPCA to the United States weather data. The fits in (a) are for four randomly selected weather stations in the dataset. The plots in (b) show the estimated mean function with perturbations from each functional principal component basis function: $\widehat{\mu}(t) \pm \delta\widehat{\psi}_l(t)$, $l = 1, \ldots, 4$.

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 applications, with 365 observations for each site.

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

from various weather stations. In their application, they uncovered four functional principal component basis functions. Similarly, we conducted VMP simulations with $L = 4$. The results are presented in Figure 5. In Figure 5 (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. 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 5 (b), we show the effect of perturbing the estimated mean function with each basis function: $\widehat{\mu}(t) \pm \delta\widehat{\psi}_l(t)$, $l = 1,\ldots,4$. The plus (minus) signs indicate the shift that each basis function makes to the mean function with a positive (negative) perturbation. The first functional principal component basis function, which accounts for 99.1% of the total variation, is a mean shift that shifts the mean function in the negative (positive) direction when it is added (subtracted). This effect is stronger in the Winter months than the Summer months, indicating that US weather is most variable in the Winter. The second functional principal component basis function, which accounts 7.2% of the total variation, shifts the mean function in the positive (negative) direction in the Summer (Winter) months when it is added. Therefore, it accounts for uniformity in the measured temperatures. As a consequence, weather stations at locations with larger discrepancies between Winter and Summer temperatures will have a strong and positive score for this basis function. The third and fourth functional principal components are harder to interpret given their weak contributions to the total variation. The scores associated with the first two basis functions for the displayed weather stations are (4.46, -0.81) for the weather station in Connecticut, (-1.56, -0.04) for the weather station in Kansas, (7.81, 2.37) for the weather station in North Dakota and (-2.57, -1.13) for the weather station in Texas. The scores for the first basis functions indicate that the lowest temperatures are to be expected in Connecticut and North Dakota, whereas Texas and Kansas have warmer temperatures. Furthermore, the scores for the second basis functions show that the greatest variability between Summer and Winter months can be found in North Dakota, whereas Connecticut and Texas may appear to be more uniform in comparison to North Dakota. On the other hand, the contribution of the second basis function to the Kansas weather station is very weak.

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. FPCAGAUSSIANLIKELIHOODFRAGMENT (Algorithm 1)
2. and MEANANDFPCGAUSSIANPENALIZATIONFRAGMENT (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 basis functions 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 50 times faster than the MCMC algorithm depending on the number of response curves. An application to a large US weather 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.

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., 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.

Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, *10*, 25–37.

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

Knowles, D. A., & Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in neural information processing systems* (pp. 1701–1709).

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. (2018). Variational message passing for skew *t* regression. *Stat*, *7*, e196.

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

McLean, M. W., & Wand, M. P. (2019). Variational message passing for elaborate response regression models. *Bayesian Analysis*, *14*, 371–398.

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.

Monahan, J. F., & Stefanski, L. A. (1992). Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral. In *Handbook of the Logistic Distribution* (pp. 529–540).

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.

Nolan, T. H., & Wand, M. P. (2017). Accurate logistic variational message passing: Algebraic and numerical details. *Stat*, *6*, 102–112.

Nolan, T. H., & Wand, M. P. (2020). Streamlined solutions to multilevel sparse matrix problems. *ANZIAM Journal*, *62*, 18–41.

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

Parisi, G. (1988). *Statistical Field Theory*. Redwood City, CA: Addison-Wesley.

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/`

Rohde, D., & Wand, M. P. (2016). Semiparametric mean field variational Bayes: General principles and numerical issues. *Journal of Machine Learning Research*, *17*, 1–47.

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.

A Proof of Theorem 3.1

We first note that

$$y_i(t) - \mu(t) = \sum_{l=1}^{L} \zeta_{il}\psi_l(t), \quad i = 1,\ldots,n. \tag{56}$$

The existence of an orthonormal basis $\psi_1^*,\ldots,\psi_L^*$ can be established via Gram-Schmidt orthogonalization. We first set

$$\phi_1 \equiv \psi_1, \quad \phi_j \equiv \psi_j - \sum_{l=1}^{j-1} \frac{\langle \phi_l, \psi_j \rangle}{||\phi_l||^2}\phi_l, \quad j = 2,\ldots,L.$$

Next, set

$$\phi_j^* = \frac{\phi_j}{||\phi_j||}, \quad j = 1,\ldots,L.$$

Then $\phi_1^*,\ldots,\phi_L^*$ form an orthonormal basis for the span of $\psi_1,\ldots,\psi_L$. Therefore, (56) can be rewritten as

$$y_i(t) - \mu(t) = \sum_{l=1}^{L} \upsilon_{il}\phi_l^*(t), \quad i = 1,\ldots,n,$$

where

$$\iota_{il} \equiv \zeta_{il}||\phi_l|| + \sum_{j=l+1}^{L} \zeta_{ij} \frac{\langle \phi_l, \psi_j \rangle}{||\phi_l||}, \quad l = 1, \ldots, L-1, \quad \iota_{iL} \equiv \zeta_{iL}||\phi_L||.$$

Note that $\iota_{i1}, \ldots, \iota_{iL}$ are correlated.

Now, define $\boldsymbol{\iota}_i \equiv (\iota_{i1}, \ldots, \iota_{iL})^\intercal$, $i = 1, \ldots, n$. Since the curves $y_1, \ldots, y_n$ are random observations of a Gaussian process, we have

$$\boldsymbol{\iota}_i \overset{\text{ind.}}{\sim} \mathrm{N}(0, \boldsymbol{\Sigma}_\iota), \quad i = 1, \ldots, n.$$

Next, establish the eigendecomposition of $\boldsymbol{\Sigma}_\iota$, such that $\boldsymbol{\Sigma}_\iota = \boldsymbol{Q}_\iota \boldsymbol{\Lambda}_\iota \boldsymbol{Q}_\iota^\intercal$, where $\boldsymbol{\Lambda}_\iota$ is a diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Sigma}_\iota$ in descending order, and the columns of $\boldsymbol{Q}_\iota$ are the corresponding eigenvectors. Then, it can be easily seen that

$$\boldsymbol{\zeta}_i^* \equiv \boldsymbol{Q}_\iota^\intercal \boldsymbol{\iota}_i \overset{\text{ind.}}{\sim} \mathrm{N}(0, \boldsymbol{\Lambda}_\iota), \quad i = 1, \ldots, n.$$

That is, the elements of $\boldsymbol{\zeta}_i^*$ are uncorrelated and $\boldsymbol{\zeta}_1^*, \ldots, \boldsymbol{\zeta}_n^*$ are independent.

Next, define the eigenvectors of $\boldsymbol{\Sigma}_\iota$ as $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_L$, such that $\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_L \end{bmatrix}$. Furthermore, define the elements of each of the eigenvectors such that $\boldsymbol{q}_l = (q_{l1}, \ldots, q_{lL})^\intercal$, $l = 1, \ldots, L$. Then, set

$$\psi_l^* \equiv \sum_{j=1}^{L} q_{lj} \phi_j^*, \quad l = 1, \ldots, L.$$

The orthonormality of $\psi_1^*, \ldots, \psi_L^*$ is easily verified:

$$\langle \psi_l^*, \psi_j^* \rangle = \left\langle \sum_{m=1}^{L} q_{lm} \phi_m^*, \sum_{k=1}^{L} q_{jk} \phi_k^* \right\rangle = \sum_{m=1}^{L} \sum_{k=1}^{L} q_{lm} q_{jk} \langle \phi_m^*, \phi_k^* \rangle = \sum_{k=1}^{L} q_{lk} q_{jk} = \boldsymbol{q}_l^\intercal \boldsymbol{q}_j = \mathbb{I}(l = j),$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Finally, we have

$$y_i(t) - \mu(t) = \sum_{l=1}^{L} \iota_{il} \phi_l^*(t) = \sum_{l=1}^{L} \sum_{j=1}^{L} \zeta_{ij}^* q_{jl} \phi_l^*(t) = \sum_{j=1}^{L} \zeta_{ij}^* \sum_{l=1}^{L} q_{jl} \phi_l^*(t) = \sum_{j=1}^{L} \zeta_{ij}^* \, \psi_j^*(t).$$

Assumptions 1 and 2 ensure that this decomposition is unique.

B Proof of Lemma 4.1

To prove (51), we first note that posterior curve estimates from the VMP algorithm satisfy

$$
\begin{aligned}
\widehat{y}_i(\boldsymbol{t}_g) &= \boldsymbol{C}_g\,\mathbb{E}_q(\boldsymbol{\nu}_\mu) + \sum_{l=1}^{L} \mathbb{E}_q(\zeta_{il})\boldsymbol{C}_g\,\mathbb{E}_q(\boldsymbol{\nu}_{\psi_l}) \\
&= \mathbb{E}_q\{\mu(\boldsymbol{t}_g)\} + \sum_{l=1}^{L} \mathbb{E}_q(\zeta_{il})\,\mathbb{E}_q\{\psi_l(\boldsymbol{t}_g)\} \\
&= \mathbb{E}_q\{\mu(\boldsymbol{t}_g)\} + \boldsymbol{\Psi}\,\mathbb{E}_q(\boldsymbol{\zeta}_i) \\
&= \mathbb{E}_q\{\mu(\boldsymbol{t}_g)\} + \boldsymbol{U}_\psi \boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal\,\mathbb{E}_q(\boldsymbol{\zeta}_i) \\
&= [\mathbb{E}_q\{\mu(\boldsymbol{t}_g)\} + \boldsymbol{U}_\psi \boldsymbol{m}_\zeta] + \boldsymbol{U}_\psi\{\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal\,\mathbb{E}_q(\boldsymbol{\zeta}_i) - \boldsymbol{m}_\zeta\} \\
&= \widehat{\mu}(\boldsymbol{t}_g) + \boldsymbol{U}_\psi \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^\intercal\{\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal\,\mathbb{E}_q(\boldsymbol{\zeta}_i) - \boldsymbol{m}_\zeta\} \\
&= \widehat{\mu}(\boldsymbol{t}_g) + \widetilde{\boldsymbol{\Psi}}\widetilde{\boldsymbol{\zeta}}_i,
\end{aligned}
\tag{57}
$$

where $\widetilde{\boldsymbol{\zeta}}_i \equiv (\widetilde{\zeta}_{i1}, \ldots, \widetilde{\zeta}_{iL})^\intercal$, $i = 1, \ldots, n$. Next, define

$$
\boldsymbol{Y} \equiv \begin{bmatrix} \widehat{y}_1(\boldsymbol{t}_g) & \ldots & \widehat{y}_n(\boldsymbol{t}_g) \end{bmatrix}
$$

Then, (57) implies

$$
\boldsymbol{Y} - \mu^*(\boldsymbol{t}_g)\boldsymbol{1}_N^\intercal = \widetilde{\boldsymbol{\Psi}}\widetilde{\boldsymbol{\Xi}}^\intercal.
$$

Now, let $\boldsymbol{c}$ be the $L \times 1$ vector, with $||\widetilde{\psi}_l||$ as the lth entry, $l = 1, \ldots, L$. Furthermore, let $1/\boldsymbol{c}$ be the $L \times 1$ vector, with $1/||\widetilde{\psi}_l||$ as the lth entry, $l = 1, \ldots, L$. Recall that we can approximate these values through numerical integration. Then,

$$
\boldsymbol{Y} - \mu^*(\boldsymbol{t}_g)\boldsymbol{1}_N^\intercal = \widetilde{\boldsymbol{\Psi}}\,\mathrm{diag}(1/\boldsymbol{c})\,\mathrm{diag}(\boldsymbol{c})\widetilde{\boldsymbol{\Xi}}^\intercal.
$$

It is easy to see that this implies (51).

C Proof of Proposition 4.2

The independence of $\widehat{\boldsymbol{\zeta}}_1, \ldots, \widehat{\boldsymbol{\zeta}}_n$ is a consequence of the independence assumption in (25). Let $\boldsymbol{c}$ and $1/\boldsymbol{c}$ retain their definitions from Appendix B. Then, note that

$$
\widehat{\boldsymbol{\zeta}}_i = \mathrm{diag}(\boldsymbol{c})\widetilde{\boldsymbol{\zeta}}_i = \mathrm{diag}(\boldsymbol{c})\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^\intercal\{\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal\,\mathbb{E}_q(\boldsymbol{\zeta}_i) - \boldsymbol{m}_\zeta\}.
$$

Recall that $\boldsymbol{m}_\zeta$ is the mean vector of the columns of $\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal \boldsymbol{\Xi}^\intercal$. Then, it is easy to see that $\sum_{i=1}^n \widehat{\boldsymbol{\zeta}}_i = \boldsymbol{0}$. Next,

$$
\begin{aligned}
\sum_{i=1}^{n} \widehat{\boldsymbol{\zeta}}_i\widehat{\boldsymbol{\zeta}}_i^\intercal &= \mathrm{diag}(\boldsymbol{c})\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^\intercal \sum_{i=1}^{n} \left[\{\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal\,\mathbb{E}_q(\boldsymbol{\zeta}_i) - \boldsymbol{m}_\zeta\}\{\boldsymbol{D}_\psi \boldsymbol{V}_\psi^\intercal\,\mathbb{E}_q(\boldsymbol{\zeta}_i) - \boldsymbol{m}_\zeta\}^\intercal\right] \boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}\,\mathrm{diag}(\boldsymbol{c}) \\
&= (n-1)\,\mathrm{diag}(\boldsymbol{c})\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^\intercal \boldsymbol{C}_\zeta \boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}\,\mathrm{diag}(\boldsymbol{c})
\end{aligned}
$$

$$= (n-1)\,\mathrm{diag}(\boldsymbol{c})\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}\,\mathrm{diag}(\boldsymbol{c})$$
$$= (n-1)\,\mathrm{diag}(\boldsymbol{c}^2),$$

which proves the results for the estimated scores.

From Lemma 4.1, we have

$$\sum_{i=1}^{n}\widehat{y}_i(\boldsymbol{t}_g) = \sum_{i=1}^{n}\left\{\widehat{\mu}(\boldsymbol{t}_g) + \sum_{l=1}^{L}\widehat{\zeta}_{il}\widehat{\psi}_l(\boldsymbol{t}_g)\right\} = \sum_{i=1}^{n}\left\{\widehat{\mu}(\boldsymbol{t}_g) + \widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\zeta}}_i\right\} = n\widehat{\mu}(\boldsymbol{t}_g),$$

where $\widehat{\boldsymbol{\Psi}} \equiv \left[\widehat{\psi}_1(\boldsymbol{t}_g)\ \cdots\ \widehat{\psi}_L(\boldsymbol{t}_g)\right]$. Therefore, the sample covariance matrix of $\widehat{y}_1(\boldsymbol{t}_g),\ldots,\widehat{y}_n(\boldsymbol{t}_g)$ is such that

$$\begin{aligned}
&\sum_{i=1}^{n}\left[\widehat{y}_i(\boldsymbol{t}_g) - \widehat{\mu}(\boldsymbol{t}_g)\right]\left[\widehat{y}_i(\boldsymbol{t}_g) - \widehat{\mu}(\boldsymbol{t}_g)\right]^{\mathsf{T}} \\
&\qquad = \sum_{i=1}^{n}\left(\sum_{l=1}^{L}\widehat{\zeta}_{il}\widehat{\psi}_l(\boldsymbol{t}_g)\right)\left(\sum_{l=1}^{L}\widehat{\zeta}_{il}\widehat{\psi}_l(\boldsymbol{t}_g)\right)^{\mathsf{T}} \\
&\qquad = \sum_{i=1}^{n}\left(\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\zeta}}_i\right)\left(\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\zeta}}_i\right)^{\mathsf{T}} \\
&\qquad = \widehat{\boldsymbol{\Psi}}\left\{\sum_{i=1}^{n}\left(\widehat{\boldsymbol{\zeta}}_i\widehat{\boldsymbol{\zeta}}_i^{*\mathsf{T}}\right)\right\}\widehat{\boldsymbol{\Psi}}^{\mathsf{T}} \\
&\qquad = (n-1)\widehat{\boldsymbol{\Psi}}\,\mathrm{diag}(\boldsymbol{c}^2)\widehat{\boldsymbol{\Psi}}^{\mathsf{T}}.
\end{aligned}$$

Simple rearrangement confirms that this is the eigenvalue decomposition of the sample covariance matrix of $\widehat{y}_1(\boldsymbol{t}_g),\ldots,\widehat{y}_n(\boldsymbol{t}_g)$, proving the result for the vectors $\widehat{\psi}_1(\boldsymbol{t}_g),\ldots,\widehat{\psi}_L(\boldsymbol{t}_g)$.

D Derivation of the FPCA Gaussian Likelihood Fragment

From (28), we have, for $i = 1,\ldots,n$,

$$\log p(\boldsymbol{y}_i|\boldsymbol{\nu},\boldsymbol{\zeta}_i,\sigma_\varepsilon^2) = -\frac{T_i}{2}\log(\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\left|\left|\boldsymbol{y}_i - \boldsymbol{C}_i\left(\boldsymbol{\nu}_\mu - \sum_{l=1}^{L}\zeta_{il}\boldsymbol{\nu}_{\psi_l}\right)\right|\right|^2 + \text{const.} \tag{58}$$

First, we establish the natural parameter vector for each of the optimal posterior density functions. These natural parameter vectors are essential for determining expectations with respect to the optimal posterior distribution. According to (10), the natural parameter vector for $q(\boldsymbol{\nu})$ is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\boldsymbol{\nu}} = \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\rightarrow\boldsymbol{\nu}} + \boldsymbol{\eta}_{\boldsymbol{\nu}\rightarrow p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)},$$

the natural parameter vector for $q(\boldsymbol{\zeta}_i)$, $i = 1,\ldots,n$, is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\boldsymbol{\zeta}_i} = \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\rightarrow\boldsymbol{\zeta}_i} + \boldsymbol{\eta}_{\boldsymbol{\zeta}_i\rightarrow p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)},$$

and the natural parameter vector for $q(\sigma_\varepsilon^2)$ is

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\sigma_\varepsilon^2} = \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\rightarrow\sigma_\varepsilon^2} + \boldsymbol{\eta}_{\sigma_\varepsilon^2\rightarrow p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)}.$$

Next, we consider the updates for standard expectations that occur for each of the random variables and random vectors in (58). For $\boldsymbol{\nu}$, we need to determine the mean vector $\mathbb{E}_q(\boldsymbol{\nu})$ and the covariance matrix $\mathbb{C}\text{ov}_q(\boldsymbol{\nu})$. The expectations are taken with respect to the normalization of

$$m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\rightarrow\boldsymbol{\nu}}(\boldsymbol{\nu})m_{\boldsymbol{\nu}\rightarrow p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)}(\boldsymbol{\nu}),$$

which is a Multivariate Normal density function with natural parameter vector $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\boldsymbol{\nu}}$. From (12), we have

$$\begin{aligned} \mathbb{E}_q(\boldsymbol{\nu}) &\longleftarrow -\frac{1}{2}\left[\text{vec}^{-1}\left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\boldsymbol{\nu}}\right)_2\right\}\right]^{-1}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\boldsymbol{\nu}}\right)_1 \\ \text{and}\quad \mathbb{C}\text{ov}_q(\boldsymbol{\nu}) &\longleftarrow -\frac{1}{2}\left[\text{vec}^{-1}\left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\boldsymbol{\nu}}\right)_2\right\}\right]^{-1}. \end{aligned} \tag{59}$$

Furthermore, the mean vector has the form

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \left\{\mathbb{E}_q(\boldsymbol{\nu}_\mu)^\intercal, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_1})^\intercal, \ldots, \mathbb{E}_q(\boldsymbol{\nu}_{\psi_L})^\intercal\right\}^\intercal, \tag{60}$$

and the covariance matrix has the form

$$\mathbb{C}\text{ov}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_\mu) & \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_\mu,\boldsymbol{\nu}_{\psi_1}) & \ldots & \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_\mu,\boldsymbol{\nu}_{\psi_L}) \\ \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\psi_1},\boldsymbol{\nu}_\mu) & \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\psi_1}) & \ldots & \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\psi_1},\boldsymbol{\nu}_{\psi_L}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\psi_L},\boldsymbol{\nu}_\mu) & \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\psi_L},\boldsymbol{\nu}_{\psi_1}) & \ldots & \mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\psi_L}) \end{bmatrix}. \tag{61}$$

Similarly, for each $i = 1,\ldots,n$, we need to determine the optimal mean vector and covariance matrix for ζ_i, which are $\mathbb{E}_q(\zeta_i)$ and $\mathbb{C}\text{ov}_q(\zeta_i)$, respectively. The expectations are taken with respect to the normalization of

$$m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\rightarrow\zeta_i}(\zeta_i)m_{\zeta_i\rightarrow p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)}(\zeta_i),$$

which is a Multivariate Normal density function with natural parameter vector $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\zeta_i}$. According to (13),

$$\begin{aligned} \mathbb{E}_q(\zeta_i) &\longleftarrow -\frac{1}{2}\left[\text{vec}^{-1}\left\{\boldsymbol{D}_L^{+\intercal}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\zeta_i}\right)_2\right\}\right]^{-1}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\zeta_i}\right)_1 \\ \text{and}\quad \mathbb{C}\text{ov}_q(\zeta_i) &\longleftarrow -\frac{1}{2}\left[\text{vec}^{-1}\left\{\boldsymbol{D}_L^{+\intercal}\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\zeta_i}\right)_2\right\}\right]^{-1}, \quad \text{for } i = 1,\ldots,n. \end{aligned} \tag{62}$$

Finally, for σ_ε^2, we need to determine $\mathbb{E}_q(1/\sigma_\varepsilon^2)$, with the expectation taken with respect to the normalization of

$$m_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\to\sigma_\varepsilon^2}(\sigma_\varepsilon^2)\, m_{\sigma_\varepsilon^2\to p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)}(\sigma_\varepsilon^2).$$

This is an Inverse $-\chi^2$ density function, with natural parameter vector $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\sigma_\varepsilon^2}$. According to Result 6 of Maestrini and Wand (2020),

$$\mathbb{E}_q(1/\sigma_\varepsilon^2) \longleftarrow \frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\sigma_\varepsilon^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)\leftrightarrow\sigma_\varepsilon^2}\right)_2}.$$

Now, we turn our attention to the derivation of the message passed from $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)$ to $\boldsymbol{\nu}$. Notice that

$$\boldsymbol{C}_i\left(\boldsymbol{\nu}_\mu - \sum_{l=1}^{L}\zeta_{il}\boldsymbol{\nu}_{\psi_l}\right) = (\widetilde{\boldsymbol{\zeta}}_i^\mathsf{T}\otimes\boldsymbol{C}_i)\boldsymbol{\nu}. \tag{63}$$

Therefore, as a function of $\boldsymbol{\nu}$, (58) can be re-written as

$$\begin{aligned}\log p(\boldsymbol{y}_i|\boldsymbol{\nu},\zeta_i,\sigma_\varepsilon^2) &= -\frac{1}{2\sigma_\varepsilon^2}\left|\left|\boldsymbol{y}_i - (\widetilde{\boldsymbol{\zeta}}_i^\mathsf{T}\otimes\boldsymbol{C}_i)\boldsymbol{\nu}\right|\right|^2 + \text{terms not involving } \boldsymbol{\nu}\\ &= \begin{bmatrix}\boldsymbol{\nu}\\ \text{vec}(\boldsymbol{\nu}\boldsymbol{\nu}^\mathsf{T})\end{bmatrix}^\mathsf{T}\begin{bmatrix}\frac{1}{\sigma_\varepsilon^2}(\widetilde{\boldsymbol{\zeta}}_i^\mathsf{T}\otimes\boldsymbol{C}_i)^\mathsf{T}\boldsymbol{y}_i\\ -\frac{1}{2\sigma_\varepsilon^2}\,\text{vec}\left\{(\widetilde{\boldsymbol{\zeta}}_i\widetilde{\boldsymbol{\zeta}}_i^\mathsf{T})\otimes(\boldsymbol{C}_i^\mathsf{T}\boldsymbol{C}_i)\right\}\end{bmatrix} + \text{terms not involving } \boldsymbol{\nu}.\end{aligned}$$

where, for each $i=1,\ldots,n$, $\widetilde{\boldsymbol{\zeta}}_i$ is defined in (31). From (6) and (27), the message from the factor $p(\boldsymbol{y}|\boldsymbol{\nu},\zeta_1,\ldots,\zeta_n,\sigma_\varepsilon^2)$ to $\boldsymbol{\nu}$ is as given in (29), which is proportional to a Multivariate Normal density function. The update for the message's natural parameter vector, in (30), is dependent upon the mean vector and covariance matrix of $\widetilde{\boldsymbol{\zeta}}_i$, which are

$$\mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i) = \{1,\mathbb{E}_q(\boldsymbol{\zeta}_i)^\mathsf{T}\}^\mathsf{T} \quad\text{and}\quad \mathbb{C}\text{ov}_q(\widetilde{\boldsymbol{\zeta}}_i) = \text{blockdiag}\left\{0,\mathbb{C}\text{ov}_q(\boldsymbol{\zeta}_i)\right\}, \quad \text{for } i=1,\ldots,n, \tag{64}$$

where $\mathbb{E}_q(\boldsymbol{\zeta}_i)$ and $\mathbb{C}\text{ov}_q(\boldsymbol{\zeta}_i)$ are defined in (62). Note that a standard statistical result allows us to write

$$\mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i\widetilde{\boldsymbol{\zeta}}_i^\mathsf{T}) = \mathbb{C}\text{ov}_q(\widetilde{\boldsymbol{\zeta}}_i) + \mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i)\,\mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i)^\mathsf{T}, \quad \text{for } i=1,\ldots,n. \tag{65}$$

Next, notice that

$$\sum_{l=1}^{L}\zeta_{il}\boldsymbol{\nu}_{\psi_l} = \boldsymbol{V}_\psi\boldsymbol{\zeta}_i, \tag{66}$$

where $\boldsymbol{V}_\psi$ is defined in (34). Then, for each $i=1,\ldots,n$, the log-density function in (58) can be represented as a function of $\boldsymbol{\zeta}_i$ by

$$\log p(\boldsymbol{y}_i|\boldsymbol{\nu},\boldsymbol{\zeta}_i,\sigma^2_\varepsilon) = -\frac{1}{2\sigma^2_\varepsilon}\left|\left|\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{\nu}_\mu - \boldsymbol{C}_i\boldsymbol{V}_\psi\boldsymbol{\zeta}_i\right|\right|^2 + \text{terms not involving } \boldsymbol{\zeta}_i$$
$$= \begin{bmatrix} \boldsymbol{\zeta}_i \\ \text{vech}(\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal) \end{bmatrix}^\intercal \begin{bmatrix} \frac{1}{\sigma^2_\varepsilon}(\boldsymbol{V}_\psi^\intercal\boldsymbol{C}_i^\intercal\boldsymbol{y}_i - \boldsymbol{h}_{\mu\psi,i}) \\ -\frac{1}{2\sigma^2_\varepsilon}\boldsymbol{D}_L^\intercal\,\text{vec}(\boldsymbol{H}_{\psi,i}) \end{bmatrix} + \text{terms not involving } \boldsymbol{\zeta}_i,$$

where $\boldsymbol{h}_{\mu\psi,i}$ and $\boldsymbol{H}_{\psi,i}$ are also defined in (34). From (6) and (27), the message from the factor $p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)$ to $\boldsymbol{\zeta}_i$ is as given in (32), which is proportional to a Multivariate Normal density function. The message's natural parameter vector update, in (33), is dependant on the following expectations that are yet to be determined:

$$\mathbb{E}_q(\boldsymbol{V}_\psi) \quad \text{and} \quad \mathbb{E}_q(\boldsymbol{H}_{\psi,i}), \quad \mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i}), \quad i = 1,\ldots,n.$$

Now, from (34),

$$\mathbb{E}_q(\boldsymbol{V}_\psi) = \left[\mathbb{E}_q(\boldsymbol{\nu}_{\psi_1}) \quad \ldots \quad \mathbb{E}_q(\boldsymbol{\nu}_{\psi_L})\right], \tag{67}$$

where, for $l = 1,\ldots,L$, $\mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})$ is defined by (59) and (60). Next, $\mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i})$ is an $L\times 1$ vector, with lth component being

$$\mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i})_l = \text{tr}\{\mathbb{C}\text{ov}_q(\boldsymbol{\nu}_\mu,\boldsymbol{\nu}_{\psi_l})\boldsymbol{C}_i^\intercal\boldsymbol{C}_i\} + \mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})^\intercal\boldsymbol{C}_i^\intercal\boldsymbol{C}_i\mathbb{E}_q(\boldsymbol{\nu}_\mu), \quad l = 1,\ldots,L, \tag{68}$$

which depends on sub-vectors of $\mathbb{E}_q(\boldsymbol{\nu})$ and sub-blocks of $\mathbb{C}\text{ov}_q(\boldsymbol{\nu})$ that are defined in (60) and (61), respectively. Finally, $\mathbb{E}_q(\boldsymbol{H}_{\psi,i})$ is an $L\times L$ matrix, with (l,l') component being

$$\mathbb{E}_q(\boldsymbol{H}_{\psi,i})_{l,l'} = \text{tr}\{\mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\psi_{l'}},\boldsymbol{\nu}_{\psi_l})\boldsymbol{C}_i^\intercal\boldsymbol{C}_i\} + \mathbb{E}_q(\boldsymbol{\nu}_{\psi_l})^\intercal\boldsymbol{C}_i^\intercal\boldsymbol{C}_i\mathbb{E}_q(\boldsymbol{\nu}_{\psi_{l'}}), \quad 1 \le l,l' \le L. \tag{69}$$

The final message to consider is the message from $p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)$ to σ^2_ε. As a function of σ^2_ε, (58) takes the form

$$\log p(\boldsymbol{y}_i|\boldsymbol{\nu},\boldsymbol{\zeta}_i,\sigma^2_\varepsilon) = -\frac{T_i}{2}\log(\sigma^2_\varepsilon) - \frac{1}{2\sigma^2_\varepsilon}\left|\left|\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{V}\widetilde{\boldsymbol{\zeta}}_i\right|\right|^2 + \text{terms not involving } \sigma^2_\varepsilon$$
$$= \begin{bmatrix} \log(\sigma^2_\varepsilon) \\ \frac{1}{\sigma^2_\varepsilon} \end{bmatrix}^\intercal \begin{bmatrix} -\frac{T_i}{2} \\ -\frac{1}{2}\left|\left|\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{V}\widetilde{\boldsymbol{\zeta}}_i\right|\right|^2 \end{bmatrix} + \text{terms not involving } \sigma^2_\varepsilon,$$

where $\boldsymbol{V}$ is defined in (37) and, for each $i = 1,\ldots,n$, $\widetilde{\boldsymbol{\zeta}}_i$ is defined in (31). From (6) and (27), the message from $p(\boldsymbol{y}|\boldsymbol{\nu},\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n,\sigma^2_\varepsilon)$ to σ^2_ε is as given in (35), which is proportional to a Inverse $-\chi^2$ density function. The message's natural parameter vector, in (36), depends on the mean of the square norm $||\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{V}\widetilde{\boldsymbol{\zeta}}_i||^2$, for $i = 1,\ldots,n$. This expectation takes the form

$$\mathbb{E}_q\left(\left|\left|\boldsymbol{y}_i - \boldsymbol{C}_i\boldsymbol{V}\widetilde{\boldsymbol{\zeta}}_i\right|\right|^2\right) = \boldsymbol{y}_i^\intercal\boldsymbol{y}_i - 2\,\mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i)^\intercal\,\mathbb{E}_q(\boldsymbol{V})^\intercal\boldsymbol{C}_i^\intercal\boldsymbol{y}_i + \operatorname{tr}\left[\left\{\mathbb{C}\mathrm{ov}_q(\widetilde{\boldsymbol{\zeta}}_i) + \mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i)\,\mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i)^\intercal\right\}\mathbb{E}_q(\boldsymbol{H}_i)\right],$$

where we introduce the matrices

$$\boldsymbol{H}_i \equiv \begin{bmatrix} h_{\mu,i} & \boldsymbol{h}_{\mu\psi,i}^\intercal \\ \boldsymbol{h}_{\mu\psi,i} & \boldsymbol{H}_{\psi,i} \end{bmatrix}, \quad \text{for } i = 1,\ldots,n, \tag{70}$$

and vectors

$$h_{\mu,i} \equiv \boldsymbol{\nu}_\mu^\intercal\boldsymbol{C}_i\boldsymbol{C}_i\boldsymbol{\nu}_\mu, \quad \text{for } i = 1,\ldots,n. \tag{71}$$

For each $i = 1,\ldots,n$, the mean vector $\mathbb{E}_q(\widetilde{\boldsymbol{\zeta}}_i)$ and $\mathbb{C}\mathrm{ov}_q(\widetilde{\boldsymbol{\zeta}}_i)$ are defined in (64). However, $\mathbb{E}_q(\boldsymbol{V})$ and $\mathbb{E}_q(\boldsymbol{H}_i)$, $i = 1,\ldots,n$, are yet to be determined. From (37),

$$\mathbb{E}_q(\boldsymbol{V}) = \left[\mathbb{E}_q(\boldsymbol{\nu}_\mu) \quad \mathbb{E}_q(\boldsymbol{\nu}_{\psi_1}) \quad \ldots \quad \mathbb{E}_q(\boldsymbol{\nu}_{\psi_L})\right],$$

where the component mean vectors are defined by (60). For each $i = 1,\ldots,n$, the expectation of $\boldsymbol{H}_i$, defined in (70), with respect to the optimal posterior distribution is

$$\mathbb{E}_q(\boldsymbol{H}_i) \equiv \begin{bmatrix} \mathbb{E}_q(h_{\mu,i}) & \mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i})^\intercal \\ \mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i}) & \mathbb{E}_q(\boldsymbol{H}_{\psi,i}) \end{bmatrix},$$

where $h_{\mu,i}$ is defined in (71) with expected value

$$\mathbb{E}_q(h_{\mu,i}) \equiv \operatorname{tr}\{\mathbb{C}\mathrm{ov}_q(\boldsymbol{\nu}_\mu)\boldsymbol{C}_i^\intercal\boldsymbol{C}_i\} + \mathbb{E}_q(\boldsymbol{\nu}_\mu)^\intercal\boldsymbol{C}_i^\intercal\boldsymbol{C}_i\,\mathbb{E}_q(\boldsymbol{\nu}_\mu).$$

Furthermore, $\mathbb{E}_q(\boldsymbol{h}_{\mu\psi,i})$ and $\mathbb{E}_q(\boldsymbol{H}_{\psi,i})$ are defined in (68) and (69), respectively.

The FPCA Gaussian likelihood fragment, summarized in Algorithm 1, is a proceduralization of these results.

E Derivation of the Mean and FPC Gaussian Penalization Fragment

From (39), we have, for $l = 1,\ldots,L$,

$$\begin{aligned}\log p(\boldsymbol{\nu}_\mu, \boldsymbol{\nu}_{\psi_l}|\sigma_\mu^2, \sigma_{\psi_l}^2) = &-\frac{K}{2}\log(\sigma_\mu^2) - \frac{K}{2}\log(\sigma_{\psi_l}^2) - \frac{1}{2}(\boldsymbol{\beta}_\mu - \boldsymbol{\mu}_{\boldsymbol{\beta}_\mu})^\intercal\boldsymbol{\Sigma}_{\boldsymbol{\beta}_\mu}^{-1}(\boldsymbol{\beta}_\mu - \boldsymbol{\mu}_{\boldsymbol{\beta}_\mu}) \\ &- \frac{1}{2\sigma_\mu^2}\boldsymbol{u}_\mu^\intercal\boldsymbol{u}_\mu - \frac{1}{2}(\boldsymbol{\beta}_{\psi_l} - \boldsymbol{\mu}_{\boldsymbol{\beta}_{\psi_l}})^\intercal\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\psi_l}}^{-1}(\boldsymbol{\beta}_{\psi_l} - \boldsymbol{\mu}_{\boldsymbol{\beta}_{\psi_l}}) - \frac{1}{2\sigma_{\psi_l}^2}\boldsymbol{u}_{\psi_l}^\intercal\boldsymbol{u}_{\psi_l}.\end{aligned} \tag{72}$$

First, we establish the natural parameter vector for each of the optimal posterior density functions.

As explained in Appendix D, these natural parameter vectors are essential for determining expectations with respect to the optimal posterior distribution. According to (10), the natural parameter vector for $q(\boldsymbol{\nu})$ is

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\boldsymbol{\nu}} = \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\rightarrow\boldsymbol{\nu}} + \boldsymbol{\eta}_{\boldsymbol{\nu}\rightarrow p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)},$$

the natural parameter vector for $q(\sigma_{\mu}^2)$ is

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\sigma_{\mu}^2} = \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\rightarrow\sigma_{\mu}^2} + \boldsymbol{\eta}_{\sigma_{\mu}^2\rightarrow p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)},$$

and, for $l = 1,\ldots,L$, the natural parameter vector for $q(\sigma_{\psi_l}^2)$ is

$$\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\sigma_{\psi_l}^2} = \boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\rightarrow\sigma_{\psi_l}^2} + \boldsymbol{\eta}_{\sigma_{\psi_l}^2\rightarrow p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)}.$$

Next, we consider the updates for standard expectations of each of the random variables and random vectors that appear in (72). For $\boldsymbol{\nu}$, we require the mean vector $\mathbb{E}_q(\boldsymbol{\nu})$ and covariance matrix $\mathbb{C}\text{ov}_q(\boldsymbol{\nu})$ under the optimal posterior distribution. The expectations are taken with respect to the normalization of

$$m_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\rightarrow\boldsymbol{\nu}}(\boldsymbol{\nu})m_{\boldsymbol{\nu}\rightarrow p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)}(\boldsymbol{\nu}),$$

which is a Multivariate Normal density function with natural parameter vector $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\boldsymbol{\nu}}$. From (12), we have

$$\begin{aligned} \mathbb{E}_q(\boldsymbol{\nu}) &\longleftarrow -\frac{1}{2}\left[\text{vec}^{-1}\left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\boldsymbol{\nu}}\right)_2\right\}\right]^{-1}\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\boldsymbol{\nu}}\right)_1 \\ \text{and}\quad \mathbb{C}\text{ov}_q(\boldsymbol{\nu}) &\longleftarrow -\frac{1}{2}\left[\text{vec}^{-1}\left\{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\mu}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\boldsymbol{\nu}}\right)_2\right\}\right]^{-1}. \end{aligned} \tag{73}$$

The sub-vectors and sub-matrices of $\mathbb{E}_q(\boldsymbol{\nu})$ and $\mathbb{C}\text{ov}_q(\boldsymbol{\nu})$ are identical to those in (60) and (61), respectively. For the mean and FPC Gaussian penalization fragment, however, we need to note further sub-vectors and sub-matrices. First,

$$\mathbb{E}_q(\boldsymbol{\nu}_{\mu}) \equiv \left\{\mathbb{E}_q(\boldsymbol{\beta}_{\mu})^{\mathsf{T}}, \mathbb{E}_q(\boldsymbol{u}_{\mu})^{\mathsf{T}}\right\}^{\mathsf{T}} \quad \text{and} \quad \mathbb{E}_q(\boldsymbol{\nu}_{\psi_l}) \equiv \left\{\mathbb{E}_q(\boldsymbol{\beta}_{\psi_l})^{\mathsf{T}}, \mathbb{E}_q(\boldsymbol{u}_{\psi_l})^{\mathsf{T}}\right\}^{\mathsf{T}}, \quad \text{for } l = 1,\ldots,L \tag{74}$$

and, second,

$$\mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\mu}) \equiv \begin{bmatrix} \mathbb{C}\text{ov}_q(\boldsymbol{\beta}_{\mu}) & \mathbb{C}\text{ov}_q(\boldsymbol{\beta}_{\mu},\boldsymbol{u}_{\mu}) \\ \mathbb{C}\text{ov}_q(\boldsymbol{u}_{\mu},\boldsymbol{\beta}_{\mu}) & \mathbb{C}\text{ov}_q(\boldsymbol{u}_{\mu}) \end{bmatrix} \tag{75}$$

and

$$\mathbb{C}\text{ov}_q(\boldsymbol{\nu}_{\psi_l}) \equiv \begin{bmatrix} \mathbb{C}\text{ov}_q(\boldsymbol{\beta}_{\psi_l}) & \mathbb{C}\text{ov}_q(\boldsymbol{\beta}_{\psi_l},\boldsymbol{u}_{\psi_l}) \\ \mathbb{C}\text{ov}_q(\boldsymbol{u}_{\psi_l},\boldsymbol{\beta}_{\psi_l}) & \mathbb{C}\text{ov}_q(\boldsymbol{u}_{\psi_l}) \end{bmatrix}, \quad \text{for } l = 1,\ldots,L. \tag{76}$$

For σ_μ^2, we need to determine $\mathbb{E}_q(1/\sigma_\mu^2)$, with expectation taken with respect to the normalization of

$$m_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\to\sigma_\mu^2}(\sigma_\mu^2)\, m_{\sigma_\mu^2\to p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)}(\sigma_\mu^2),$$

which is an Inverse $-\chi^2$ density function with natural parameter vector $\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\sigma_\mu^2}$. According to Result 6 of Maestrini and Wand (2020),

$$\mathbb{E}_q(1/\sigma_\mu^2) \longleftarrow \frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\sigma_\mu^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\sigma_\mu^2}\right)_2}. \tag{77}$$

Similar arguments can be used to show that

$$\mathbb{E}_q(1/\sigma_{\psi_l}^2) \longleftarrow \frac{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\psi_l}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\sigma_{\psi_l}^2}\right)_1 + 1}{\left(\boldsymbol{\eta}_{p(\boldsymbol{\nu}|\sigma_{\psi_l}^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)\leftrightarrow\sigma_{\psi_l}^2}\right)_2}, \quad \text{for } l = 1,\ldots,L. \tag{78}$$

Now, we turn our attention to the derivation of the messages passed from the factor. As a function of $\boldsymbol{\nu}$, (72) this can be re-written as

$$\begin{aligned}\log p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2) &= -\frac{1}{2}\boldsymbol{\nu}^\intercal\boldsymbol{\Sigma}_\nu^{-1}\boldsymbol{\nu} + \boldsymbol{\nu}^\intercal\boldsymbol{\Sigma}_\nu^{-1}\boldsymbol{\mu}_\nu + \text{terms not involving } \boldsymbol{\nu} \\ &= \begin{bmatrix}\boldsymbol{\nu}\\ \text{vec}(\boldsymbol{\nu}\boldsymbol{\nu}^\intercal)\end{bmatrix}^\intercal \begin{bmatrix}\boldsymbol{\Sigma}_\nu^{-1}\boldsymbol{\mu}_\nu \\ -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}_\nu^{-1})\end{bmatrix} + \text{terms not involving } \boldsymbol{\nu},\end{aligned}$$

where $\boldsymbol{\mu}_\nu$ and $\boldsymbol{\Sigma}_\nu$ are defined in (40). From (6), the message from the factor $p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)$ to $\boldsymbol{\nu}$ is as given in (41), which is proportional to a Multivariate Normal density function. The update for the message's natural parameter vector, in (42), is dependant upon the expectation of $\boldsymbol{\Sigma}_\nu^{-1}$, which is given by

$$\mathbb{E}_q(\boldsymbol{\Sigma}_\nu^{-1}) = \text{blockdiag}\left\{\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_\mu} & \mathbf{O}^\intercal \\ \mathbf{O} & \mathbb{E}_q(1/\sigma_\mu^2)\boldsymbol{I}_K\end{bmatrix}, \underset{l=1,\ldots,L}{\text{blockdiag}}\left(\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\psi_l}} & \mathbf{O}^\intercal \\ \mathbf{O} & \mathbb{E}_q(1/\sigma_{\psi_l}^2)\boldsymbol{I}_K\end{bmatrix}\right)\right\},$$

where $\mathbb{E}_q(1/\sigma_\mu^2)$ and, for $l = 1,\ldots,L$, $\mathbb{E}_q(1/\sigma_{\psi_l}^2)$ are defined in (77) and (78), respectively.

As a function of σ_μ^2, (72) can be re-written as

$$\begin{aligned}\log p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2) &= -\frac{K}{2}\log(\sigma_\mu^2) - \frac{1}{2\sigma_\mu^2}\boldsymbol{u}_\mu^\intercal\boldsymbol{u}_\mu + \text{terms not involving } \sigma_\mu^2 \\ &= \begin{bmatrix}\log(\sigma_\mu^2)\\ 1/\sigma_\mu^2\end{bmatrix}^\intercal \begin{bmatrix}-\frac{K}{2} \\ -\frac{1}{2}\boldsymbol{u}_\mu^\intercal\boldsymbol{u}_\mu\end{bmatrix} + \text{terms not involving } \sigma_\mu^2.\end{aligned}$$

From (6), the message from the factor $p(\boldsymbol{\nu}|\sigma_\mu^2,\sigma_{\psi_1}^2,\ldots,\sigma_{\psi_L}^2)$ to σ_μ^2 is as given in (43), which is an Inverse $-\chi^2$ density function upon normalization. The message's natural parameter vector update in

(44) depends on $\mathbb{E}_q(\boldsymbol{u}_\mu^\intercal \boldsymbol{u}_\mu)$. Standard statistical results and sub-vector and sub-matrix definitions in (74) and (75) can be employed to show that

$$\mathbb{E}_q(\boldsymbol{u}_\mu^\intercal \boldsymbol{u}_\mu) = \mathbb{E}_q(\boldsymbol{u}_\mu)^\intercal \mathbb{E}_q(\boldsymbol{u}_\mu) + \mathrm{tr}\left\{\mathbb{C}\mathrm{ov}_q(\boldsymbol{u}_\mu)\right\}.$$

As a function of $\sigma^2_{\psi_l}$, for $l = 1, \ldots, L$, (72) can be re-written as

$$\begin{aligned}\log p(\boldsymbol{\nu}|\sigma^2_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L}) &= -\frac{K}{2}\log(\sigma^2_{\psi_l}) - \frac{1}{2\sigma^2_{\psi_l}}\boldsymbol{u}_{\psi_l}^\intercal \boldsymbol{u}_{\psi_l} + \text{terms not involving } \sigma^2_{\psi_l} \\ &= \begin{bmatrix} \log(\sigma^2_{\psi_l}) \\ 1/\sigma^2_{\psi_l} \end{bmatrix}^\intercal \begin{bmatrix} -\frac{K}{2} \\ -\frac{1}{2}\boldsymbol{u}_{\psi_l}^\intercal \boldsymbol{u}_{\psi_l} \end{bmatrix} + \text{terms not involving } \sigma^2_{\psi_l}.\end{aligned}$$

From (6), the message from the factor $p(\boldsymbol{\nu}|\sigma^2_\mu, \sigma^2_{\psi_1}, \ldots, \sigma^2_{\psi_L})$ to $\sigma^2_{\psi_l}$ is as given in (45), which is an Inverse $-\chi^2$ density function upon normalization. The message's natural parameter vector update in (46) depends on $\mathbb{E}_q(\boldsymbol{u}_{\psi_l}^\intercal \boldsymbol{u}_{\psi_l})$. Standard statistical results and sub-vector and sub-matrix definitions in (74) and (76) can be employed to show that

$$\mathbb{E}_q(\boldsymbol{u}_{\psi_l}^\intercal \boldsymbol{u}_{\psi_l}) = \mathbb{E}_q(\boldsymbol{u}_{\psi_l})^\intercal \mathbb{E}_q(\boldsymbol{u}_{\psi_l}) + \mathrm{tr}\left\{\mathbb{C}\mathrm{ov}_q(\boldsymbol{u}_{\psi_l})\right\}.$$

The mean and FPC Gaussian penalization fragment, summarized in Algorithm 2, is a procedural-ization of these results.