Bayesian Analysis (2017) **12**, Number 2, pp. 311–333

Bayesian Estimation of Principal Components for Functional Data

Adam J. Suarez[*] and Subhashis Ghosal[†]

Abstract. The area of principal components analysis (PCA) has seen relatively few contributions from the Bayesian school of inference. In this paper, we propose a Bayesian method for PCA in the case of functional data observed with error. We suggest modeling the covariance function by use of an approximate spectral decomposition, leading to easily interpretable parameters. We perform model selection, both over the number of principal components and the number of basis functions used in the approximation. We study in depth the choice of using the implied distributions arising from the inverse Wishart prior and prove a convergence theorem for the case of an exact finite dimensional representation. We also discuss computational issues as well as the care needed in choosing hyperparameters. A simulation study is used to demonstrate competitive performance against a recent frequentist procedure, particularly in terms of the principal component estimation. Finally, we apply the method to a real dataset, where we also incorporate model selection on the dimension of the finite basis used for modeling.

Keywords: principal components, covariance estimation, functional data.

1 Introduction

In the rapidly expanding area of functional data analysis, data compression has become an oft-employed strategy. Principal component analysis (PCA) has become a widespread tool in the area of functional data, where the high dimensionality of the data can quickly become unmanageable. Principal components can be used to reconstruct a process approximately, using relatively few random variables. At its heart, PCA is an exploratory tool used to gain insight into the structure of the data. It is also used in less scrupulous endeavors, such as preprocessing for a regression analysis. For a textbook length treatment of classical multivariate PCA, see Jolliffe (1986).

PCA for functional data has likewise become a very popular technique. Ramsay and Silverman (2005) certainly helped make functional PCA (FPCA) a standard first step when dealing with functional data. For another textbook account of FPCA, see Horváth and Kokoszka (2012).

Bayesian methods for multivariate PCA have been relatively absent from the literature. Tipping and Bishop (1999) showed how the traditional method of PCA can be viewed as the solution to a maximum likelihood procedure; this likelihood was then used for a Bayesian treatment in Bishop (1999). For functional data, Behseta et al. (2005)

[*] Department of Statistics, Box 8203, North Carolina State University, Raleigh, NC 27695-8203, ajsuarez@ncsu.edu

[†] Department of Statistics, Box 8203, North Carolina State University, Raleigh, NC 27695-8203, sghosal@ncsu.edu

DOI: 10.1214/16-BA1003

proposed a Bayesian method for FPCA, and van der Linde (2008) proposed an approximate Bayesian method using variational calculations. Frequentist PCA is commonly used as the first step in multi-step procedures; one reason for the lack of the subjective perspective in the PCA literature is certainly due to the fact that Bayesian procedures are, by their nature, not performed stepwise. The extension of Bayesian procedures to more complicated situations usually comes from hierarchically integrating the simpler model into a larger one.

In this paper, we investigate a potential model for the covariance structure of functional data observed with noise. The model jointly smooths the observations and estimates the principal components. As a Bayesian procedure, model selection on the chosen number of basis functions is conceptually straightforward, and we demonstrate this on a real data set.

Section 2 provides background material. Section 3 describes the motivation for the model used for the data, along with the priors used throughout. Section 3 also discusses some issues in choosing hyperparameters for the typical priors. Section 3.6 describes the method used for model comparison in terms of the number of basis functions used for approximation. In Section 4, we prove a convergence theorem for the case of an inverse Wishart prior. Sections 5 and 6 present a simulation study and applied data example, respectively.

2 Background

Let $\{X(t) : t \in [0,1]\}$ be a stochastic process such that the sample paths of $X(t)$ are square integrable. By adopting a change of location and scale, if necessary, any bounded interval can be reduced to $[0,1]$, which we shall abbreviate by I. Usually the index represents time, and for functional observations the boundedness of the domain is most natural because data can be observed only over a limited time. Let $\mu(t) := \mathrm{E}[X(t)]$ for all $t \in I$, and let the covariance function of the process be given by $\kappa(s,t) = \mathrm{Cov}(X(s), X(t))$ for all $s,t \in I$. Note that κ is symmetric in its arguments and is a positive definite function, i.e., for any $t_1, \ldots, t_k \in I$, the matrix $(\!(\kappa(t_i,t_j))\!)$ is positive definite. We assume that κ is continuous on $I \times I$ and let $M := \sup\{|\kappa(s,t)| : s,t \in I\} < \infty$. On the space $L_2 = L_2(I)$ of square integrable functions, we use the standard inner product $\langle f, g\rangle = \int_I f(t)g(t)dt$, and the norm which this implies.

The covariance function then defines an integral operator, $T_\kappa : L_2 \to L_2$, given by

$$T_\kappa f = \int_I \kappa(\cdot, t) f(t) dt, \tag{1}$$

for all $f \in L_2$. It is well known that T_κ is a Hilbert–Schmidt operator, and, in particular, T_κ is a compact linear operator, which necessarily has a countable spectrum $\{\lambda_1, \lambda_2, \ldots\}$ with 0 as the only accumulation point (Ash, 1965). Therefore, the important quantities for us are the eigenfunctions and eigenvalues of the covariance operator, i.e., functions $\{\phi_1, \phi_2, \ldots\}$ and nonnegative real numbers $\{\lambda_1, \lambda_2, \ldots\}$ (assumed to be in nonincreasing order) such that

$$T_\kappa \phi_i = \lambda_i \phi_i. \tag{2}$$

Mercer's theorem (Ash, 1965, Theorem 2.4) states that the covariance function can be represented, for $s, t \in I$, as

$$\kappa(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t), \tag{3}$$

where the convergence is uniform on $I \times I$.

The eigenfunctions, $\{\phi_1, \phi_2, \ldots\}$ are called the principal components for reasons we will now discuss. The map $f \mapsto \mathrm{Var}\langle f, X\rangle = \mathrm{Var}(\int_I f(t)X(t)dt)$ obtains its maximum on the unit sphere at ϕ_1. The set $\{\phi_1, \phi_2, \ldots\}$ satisfies

$$\phi_i = \arg\max_{f \in L_2} \left\{ \mathrm{Var}\langle f, X\rangle : \|f\| = 1, \langle f, \phi_k\rangle = 0 \text{ for } k = 1, \ldots, i-1 \right\}. \tag{4}$$

In this way, the eigenfunctions are the principal directions of variation for the process. The other reason for the term, "principal components," is due to the Karhunen–Loève expansion theorem, which states that the process, X, as a random element of L_2, can be represented as

$$X(t) = \mu(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} Z_i \phi_i(t), \tag{5}$$

where $\{Z_1, Z_2, \ldots\}$ are uncorrelated random variables with unit variance and the convergence is in mean square, uniformly on I. In this way, the eigenfunctions can be seen as a way to decompose the process into orthogonal (uncorrelated) components. An important special case to note is that (5) holds pointwise on I almost surely if X is a Gaussian process, and then each Z_i is also a Gaussian random variable, and hence they are all independent.

3 Model, Prior Specification, and Posterior Computation

Let $X_1, X_2, \ldots$ be independent and identically distributed observations from a Gaussian process, $\mathrm{GP}(\mu, \kappa)$, on $I = [0,1]$, where μ is the mean function, and κ is the covariance function. We will assume, however, that our observations have been contaminated with additional noise, i.e., we observe noisy data on some grid of points $\{t_1, \ldots, t_T\}$. We assume that all of the data is observed on the same grid for simplicity of formulas. Let $\boldsymbol{X}_i = (X_i(t_1), \ldots, X_i(t_T))'$ be the ith underlying discretized function, and let $\boldsymbol{Y}_i = \boldsymbol{X}_i + \epsilon_i$ be the ith observation, where $\epsilon_i \overset{iid}{\sim} N_T(\mathbf{0}, \sigma^2 \boldsymbol{I}_T)$. Our goal will be to estimate κ and the principal components it induces from data, and we will do so by placing priors on all the parameters.

To put a prior on μ, we shall use a Gaussian process, which reduces to the multivariate normal distribution on the discretized observations. In putting a prior for κ, we shall construct our prior based on an approximate spectral representation by truncating the series in (3), but allowing a prior on the number of terms to ensure full support. We shall induce a prior on the eigenvalues and eigenfunctions indirectly from that on the covariance matrix on the finite grid of time points, which will be chosen as the inverse Wishart distribution.

Consider a given basis, $\{h_1.h_2,\ldots\}$, for L_2. Since any eigenfunction, ϕ_i, of T_κ can be expanded as

$$\phi_i = \sum_{j=1}^{\infty} \alpha_{ij} h_j, \tag{6}$$

a particularly convenient method of putting a prior on $\{\phi_1, \phi_2, \ldots\}$ is by truncating (6) at some level, $J \in \mathbb{N}$. Let $\boldsymbol{h}_J = (h_1, \ldots, h_J)'$. We will also truncate the expansion of κ in (3) at some level, $K \in \mathbb{N}$, putting a prior on the resulting coefficients and also a prior on J. Since the finitely truncated series converges to (6) as $J \to \infty$, this procedure ensures that the resulting objects, $\{\phi_1, \phi_2, \ldots\}$, get a fully supported prior if the coefficients get such a prior for each value of J. Furthermore, a prior on κ is induced by truncating (3) at level K, i.e. $\kappa(s,t) = \sum_{i=1}^{K} \lambda_i \phi_i(s)\phi_i(t)$, and imposing a prior on K. Let $\boldsymbol{A}_{KJ} = (\!(\alpha_{ij})\!)$ be the $K \times J$ matrix of coefficients and $\boldsymbol{\phi}_K = (\phi_1, \ldots, \phi_K)'$ be given by

$$\boldsymbol{\phi}_K = \boldsymbol{A}_{KJ} \boldsymbol{h}_J. \tag{7}$$

Let $\boldsymbol{\Lambda}_K = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$. Then the prior on κ can be induced by the relation

$$\boldsymbol{\kappa}(s,t) = \boldsymbol{h}_J'(s) \boldsymbol{A}_{KJ}' \boldsymbol{\Lambda}_K \boldsymbol{A}_{KJ} \boldsymbol{h}_J(t), \tag{8}$$

and priors on K and J. However, as mentioned above, instead of directly putting priors on $\boldsymbol{A}_{KJ}$ and $\boldsymbol{\Lambda}_K$, we proceed in the reverse order and induce priors on them through a convenient prior on $\boldsymbol{\Sigma} = \boldsymbol{A}_{KJ}' \boldsymbol{\Lambda}_K \boldsymbol{A}_{KJ}$. Details of the specification are explained below.

3.1 Model and Priors

In the following, if $K = J$, it leads to substantial simplification, although such a choice may lead to overfitting of the covariance function; it is reasonable to believe that more basis functions would be needed than the number of principal components needed to reconstruct the covariance function. We first describe the model for the case of $K = J$ to then motivate the low rank model ($K < J$). Conditional on $K = J$, the model and the prior distribution can be described by the following hierarchical scheme for $i = 1, \ldots, n$,

$$\boldsymbol{Y}_i \overset{ind}{\sim} N_T(\boldsymbol{H}_J \boldsymbol{\beta}_{i,J}, \sigma^2 \boldsymbol{I}) \tag{9a}$$

$$\boldsymbol{\beta}_{i,J} \overset{iid}{\sim} N_J(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{9b}$$

$$\boldsymbol{\theta} \sim N_J(\boldsymbol{\theta}_0, \tau \boldsymbol{\Sigma}) \tag{9c}$$

$$\sigma^2 \sim \text{inv-Gamma}(a, b) \tag{9d}$$

$$\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(\nu, \boldsymbol{\Xi}^{-1}) \tag{9e}$$

$$J \sim \text{Poisson}(j_0),\ \text{truncated to } \{1,\ldots,T-1\}, \tag{9f}$$

where $\boldsymbol{H}_J$ is a $T \times J$ matrix whose columns consist of the basis functions evaluated at all grid points. When convenient, we may drop the subscript from certain expressions.

The functions $\boldsymbol{h}(t)'\boldsymbol{\beta}_i, i = 1,\ldots,n$ correspond to the underlying (unobserved) noise-free functional observations. The function $\boldsymbol{h}(t)'\boldsymbol{\theta}$ corresponds to the overall population mean for the functional observations, $\boldsymbol{h}(s)'\boldsymbol{\Sigma}\boldsymbol{h}(t)$ is the covariance function of interest, and $\boldsymbol{A}\boldsymbol{h}(t)$ is the vector of functional principal components. When estimating the full model, that is, averaging across the posterior distribution of K, it is important to restrict attention to parameters whose dimension does not depend on K. For example, the covariance function evaluated at the observed grid points, $\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}'$, has dimension $T \times T$, and has meaning across values of K, whereas $\boldsymbol{\beta}_i$ only has meaning for a given value of K.

3.2 Low Rank Model

Because the Wishart distribution gives probability 1 to nonsingular matrices, if we wish to allow the number of principal components, K, to be less than the number of basis functions used in approximations, J, we should allow for singular Wishart matrices. The most straightforward approach is to use a singular center matrix, of rank K, in the prior specification. First, let $\Xi = \boldsymbol{U}\boldsymbol{L}\boldsymbol{U}'$, where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{L}$ is the diagonal matrix of ordered eigenvalues. Choose $\Xi_K = \boldsymbol{U}_K\boldsymbol{L}_K(\boldsymbol{U}_K)'$, where $\boldsymbol{U}_K$ is the $J \times K$ matrix formed by the first K columns of $\boldsymbol{U}$, and $\boldsymbol{L}_K$ is the $K \times K$ matrix formed by first K rows and K columns of $\boldsymbol{L}$. This implies that although Ξ_K is $J \times J$, it has rank K. Let $\boldsymbol{A}^+$ denote the Moore–Penrose inverse of a matrix $\boldsymbol{A}$. Then, if $\boldsymbol{\Omega}_1 \sim \text{Wishart}(\nu, \Xi_K^+)$ and $\boldsymbol{\Omega}_2 \sim \text{Wishart}(\nu, \boldsymbol{L}_K^{-1})$,

$$\boldsymbol{\Omega}_1 \overset{d}{=} \boldsymbol{U}_K\boldsymbol{\Omega}_2\boldsymbol{U}_K'$$
$$\boldsymbol{\Omega}_1^+ = \boldsymbol{U}_K\boldsymbol{\Omega}_2^{-1}\boldsymbol{U}_K'.$$

This gives the motivation for our full, low-rank model:

$$\boldsymbol{Y}_i \overset{ind}{\sim} N_T(\boldsymbol{H}_J\boldsymbol{U}_K\boldsymbol{\beta}_{i,K}, \sigma^2\boldsymbol{I}) \tag{10a}$$
$$\boldsymbol{\beta}_{i,K} \overset{iid}{\sim} N_K(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{10b}$$
$$\boldsymbol{\theta} \sim N_K(\boldsymbol{\theta}_0, \tau\boldsymbol{\Sigma}) \tag{10c}$$
$$\sigma^2 \sim \text{inv-Gamma}(a, b) \tag{10d}$$
$$\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(\nu, \boldsymbol{L}_K^{-1}) \tag{10e}$$
$$J \sim \text{Poisson}(j_0),\ \text{truncated to } \{1,\ldots,T-1\}, \tag{10f}$$
$$K \sim \text{Poisson}(k_0),\ \text{truncated to } \{1,\ldots,J\}. \tag{10g}$$

This model implies that $\boldsymbol{U}_K\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_K' \sim \text{Wishart}(\nu, \Xi_K^+)$ that provides the singularity we desire in our modeling. Because our prior puts mass on all (J, K) pairs with $1 \le K \le J$, we still have a marginal full-rank model, but our posterior will be mixed over rank-deficient models.

3.3 Model Choice

There is a vast literature on Bayesian model choice, and many of the possible procedures could be used for our purposes. Reversible jump Markov Chain Monte Carlo (MCMC) (Green, 1995) has become a widely used method, but presents a major hurdle in our case with the need for a proposal distribution when using the model that includes the parameter $\boldsymbol{\Sigma}$. The main challenge is respecting the positive definiteness of the matrix when proposing a jump to a higher dimension. This can be overcome by marginalizing out $\boldsymbol{\Sigma}$, obtaining posterior samples of the $\boldsymbol{\beta}_i$'s, and, finally, generating samples for $\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}'$, which would then all live in the same dimension, even though K might change between steps.

The approach we take is to estimate the posterior distribution of K through approximations to the marginal likelihood for each model. We employ independent Gibbs samplers for each value of K, and compute estimates of their marginal likelihood using results from Chib (1995). The obvious disadvantage of this strategy is the need to run MCMC chains for each possible model. The likely faster convergence of the Gibbs sampler, however, can partially offset the cost along with the ability to implement the chains in parallel. In the application considered below, the relatively small number of sampled time points lends itself well to this method, and gives much more confidence in the convergence than other approaches not based on Gibbs sampling.

3.4 Choice of Hyperparameters

Especially in the case of the inverse Wishart prior, selection of the hyperparameters is an important issue. Because of the link between principal components and covariance estimation, prior elicitation can be done in either domain. However, some problems can arise with what appears to be a good default choice. Specifically, because of the required smoothness conditions on the covariance function, it is not possible to choose the identity matrix as the center matrix for the inverse Wishart prior.

A sensible choice of $\boldsymbol{\Xi}$ (whose size depends on J) is given below. Using the prior covariance function, construct the covariance matrix corresponding to the grid being used; call this $\boldsymbol{\Sigma}^*$. We then propose using $\boldsymbol{\Xi}^{-1} = (\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{\Sigma}^*\boldsymbol{H}(\boldsymbol{H}'\boldsymbol{H})^{-1}$ as the choice of hyperparameter, which can be seen as a least-squares projection. This matrix can be shown to be invertible using the facts that $\boldsymbol{H}$ has full column rank because it is comprised of function evaluation from an orthogonal set of functions, and $\boldsymbol{\Sigma}^*$ is invertible because it is derived from a valid covariance function of a vector without linear restrictions. To complete the specification of the Wishart prior, for model (J, K), a reasonable default choice for the degrees of freedom is $\nu = 2K$, which, in the Gibbs step described later, implies the mean of the inverse Wishart distribution that resembles a typical covariance estimate with denominator $n + k - 1$. This choice is implemented in the empirical comparisons below, and performs very competitively.

Finally, for the inverse gamma prior, the choice of (a, b) is very important since it controls the amount of smoothing performed on the data. It corresponds to the prior beliefs on the amount of sampling noise present in the data. It turns out that the choice is sensitive, and, in practice, several values should be tried when performing the analysis. There is an empirical choice available for these hyperparameters. Let $\widehat{\boldsymbol{\beta}}_i = (\boldsymbol{U}'_K \boldsymbol{H}'_J \boldsymbol{H}_J \boldsymbol{U}'_K)^{-1} \boldsymbol{U}'_K \boldsymbol{H}'_J \boldsymbol{Y}_i$ be the ordinary least-squares estimate. It is well known that, conditional on σ^2 and $\{\beta_{i,K}\}_{i=1}^n$, the quantity,

$$\sigma^{-2}\mathrm{SSE}_i = \sigma^{-2}\|\boldsymbol{Y}_i - \boldsymbol{H}_J \boldsymbol{U}_K \widehat{\boldsymbol{\beta}}_i\| \overset{\text{iid}}{\sim} \chi^2_{n-K}.$$

It can be shown that the marginal distribution of SSE_i, with σ^2 integrated out, is proportional to

$$\frac{b^a}{\Gamma(a)} \left(b + \frac{1}{2} \sum_{i=1}^n \mathrm{SSE}_i \right)^{\frac{n(n-K)}{2} - n + a + 1} \Gamma\left(\frac{n(n-K)}{2} - n + a + 1 \right). \tag{11}$$

Using this, a and b can be chosen to be the values which maximize this quantity evaluated at the data.

3.5 Posterior Computation

Posterior computation is done independently for each pair (J, K), $1 \le K \le J \le J_{\max}$. The primary advantage is the ability to implement a Gibbs sampler for each model. Since the models are computed separately, they can be run in parallel to offset the computation cost of running all models. Having independent samples for each model also allows for the adjustment of the priors on J and K without the need to rerun the MCMC chains. Finally, it provides full information for each individual model in the case that a single model is desired, instead of a fully Bayesian, model-averaged posterior.

For a fixed pair (J, K), the Gibbs sampler following from the low rank model (10) is as follows:

- For $i = 1, \ldots, n$, sample $\boldsymbol{\beta}_{i,K}$ from a K-dimensional Gaussian distribution with mean

  $$\left(\sigma^{-2} \boldsymbol{U}'_K \boldsymbol{H}'_J \boldsymbol{H}_J \boldsymbol{U}_K + \boldsymbol{\Sigma}^{-1}\right)^{-1} \left(\sigma^{-1} \boldsymbol{U}'_K \boldsymbol{H}'_J \boldsymbol{Y}_i + \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right),$$

  and variance

  $$\left(\sigma^{-2} \boldsymbol{U}'_K \boldsymbol{H}'_J \boldsymbol{H}_J \boldsymbol{U}_K + \boldsymbol{\Sigma}^{-1}\right)^{-1}.$$

- Sample θ from a K-dimensional Gaussian distribution,

  $$N\left(\left(n + \tau^{-1}\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{\beta}_{i,K} + \tau^{-1}\theta_0 \right), \left(n + \tau^{-1}\right)^{-1} \boldsymbol{\Sigma} \right).$$

- Sample $\boldsymbol{\Sigma}^{-1}$ from a Wishart distribution with degrees of freedom $\nu + n + 1$, and center matrix

  $$\left(L_K + \tau^{-1} \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right)' + \sum_{i=1}^n \left(\boldsymbol{\beta}_{i,K} - \boldsymbol{\theta}\right) \left(\boldsymbol{\beta}_{i,K} - \boldsymbol{\theta}\right)' \right)^{-1}.$$

- Finally, sample σ^{-2} from a gamma distribution with shape parameter $a+n/2$ and rate parameter
$$b + \sum_{i=1} \left\| \boldsymbol{Y}_i - \boldsymbol{H}_J \boldsymbol{U}_K \boldsymbol{\beta}_{i,K} \right\|^2 .$$

3.6 Alternative Posterior Computation

Conditional on (J, K), the posterior distribution can be obtained using a Gibbs sampling scheme. When computation is performed within this paper, this was the implemented approach. Model averaging over (J, K) can be obtained by using independent MCMC chains for each value. We do wish to point out some equivalent models that have potential computational benefits. Marginalizing out $\boldsymbol{\beta}_i$ gives

$$\begin{gathered}
\boldsymbol{X}_i \overset{iid}{\sim} N_T(\boldsymbol{HU\theta}, \sigma^2 \boldsymbol{I} + \boldsymbol{HU\Sigma U'H'}) \\
\boldsymbol{\theta} \sim N_K(\boldsymbol{\theta}_0, \tau\boldsymbol{\Sigma}) \\
\sigma^2 \sim \text{inv-Gamma}(a, b) \\
\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(K, L^{-1}) \\
K \sim \text{Poisson}(k_0), \text{ truncated to } \{1, \ldots, T-1\},
\end{gathered}$$

for the top level. In this case, we see that, for the covariance to be an identifiable parameter, we require that $K < T$, so that any prior that we use should have probability one of meeting this restriction.

We can also marginalize out $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ to obtain an equivalent prior for the $\boldsymbol{\beta}_i$'s. Specifically, if we let $\boldsymbol{B}_K = [\boldsymbol{\beta}_{1,K} | \boldsymbol{\beta}_{2,K} | \cdots | \boldsymbol{\beta}_{n,K}]$, and $\boldsymbol{\Theta}_0 = [\boldsymbol{\theta}_0 | \cdots | \boldsymbol{\theta}_0]$, then the marginal prior on $\boldsymbol{B}_K$ is the so-called matrix t-distribution (Lad, 1996) with density

$$\begin{aligned}
&\left[\prod_{i=1}^{k} \frac{\Gamma\left(\frac{v+n+1-i}{2}\right)}{\Gamma\left(\frac{v+1-i}{2}\right)} \right] (\det \boldsymbol{\Xi})^{v/2} \left(\det \left(\boldsymbol{I} - (n+\tau)^{-1} \mathbf{1} \right) \right)^{k/2} \\
&\times \left(\det \left(\boldsymbol{L}_K + (\boldsymbol{B}_K - \boldsymbol{\Theta}_0) \left[\boldsymbol{I} - (n+\tau)^{-1} \mathbf{1} \right] (\boldsymbol{B}_K - \boldsymbol{\Theta}_0)' \right) \right)^{-(v+n)/2} .
\end{aligned}$$

Since we have then lost conjugacy, we can also integrate out σ^2 to represent the model in terms of $\boldsymbol{B}_K$ only. The joint density of $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ is proportional to

$$\left(1 + \frac{1}{2b} \sum_{i=1}^{n} \| \boldsymbol{Y}_i - \boldsymbol{H}_J \boldsymbol{U}_K \boldsymbol{\beta}_{i,K} \|^2 \right)^{-(a+nT/2)} . \tag{12}$$

This formulation can be useful for implementing a reversible jump MCMC scheme, so that proposals are not needed for the covariance parameter, which would require positive definiteness constraints. It can also be used to obtain the posterior mode for $\boldsymbol{\beta}_i, i = 1, \ldots, n$. This can potentially be much faster than full MCMC, especially when we would like to compare many different models. In either of these cases, we can use the following conditional posterior distributions, and expectations derived from them, to infer on other parameters:

$$\boldsymbol{\Sigma}|\boldsymbol{\beta},\boldsymbol{Y} \sim \text{inv-Wishart}\,(v+n,\boldsymbol{\Sigma}^*) \tag{13}$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{L} + \sum_{i=1}^{n}(\boldsymbol{\beta}_i-\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_i-\bar{\boldsymbol{\beta}})' + \frac{n}{n\tau+1}(\boldsymbol{\theta}_0-\bar{\boldsymbol{\beta}})(\boldsymbol{\theta}_0-\bar{\boldsymbol{\beta}})' \tag{14}$$

$$\boldsymbol{\theta}|\boldsymbol{\beta},\boldsymbol{Y},\boldsymbol{\Sigma} \sim N\left(\frac{\boldsymbol{\theta}_0+\tau\sum_{i=1}^{n}\boldsymbol{\beta}_i}{1+n\tau},(v+n)\boldsymbol{\Sigma}\right). \tag{15}$$

4 Asymptotic Results

We now study the posterior rate of contraction, ϵ_n, such that the posterior probability of the $M_n\epsilon_n$-ball around the true parameter given n observations tends to 1 as n increases to infinity for any $M_n \to \infty$.

Let Π denote the prior measure on $\mathcal{P}$, the parameter space regarded as a subset of all probability measures, with a typical element P having density p, and within which exists the true measure, P_0. For $P, Q \in \mathcal{P}$, let

$$K(p,q) = P\log\frac{p}{q}, \qquad V_+(p,q) = P\log_+^2\frac{p}{q},$$

where $\log_+ x = \max(\log x, 0)$. We will also make use of the Hellinger distance, defined as $d_H(P,Q) = (\int(\sqrt{p}-\sqrt{q})^2)^{1/2}$, and the Frobenius norm on matrices, which, for $\boldsymbol{A} = (\!(a_{ij})\!)$, is defined to be $\|\boldsymbol{A}\|_F^2 = \sum_{i,j}a_{ij}^2 = \text{tr}(\boldsymbol{A}'\boldsymbol{A})$.

Let P_0 stand for the true distribution with density p_0. In order to obtain the posterior rate of convergence, ϵ_n, we apply Theorem 2.1 of Ghosal et al. (2000). Thus, we need to verify that for a constant $C > 0$, and a sequence $\{\mathcal{P}_n\}$ of subsets of the parameter space,

$$\log N(\epsilon_n/2,\mathcal{P}_n,d) \le n\epsilon_n^2, \tag{16}$$

$$\Pi(\mathcal{P}_n^c) \le \exp\left(-n\epsilon_n^2(C+4)\right), \tag{17}$$

$$\Pi\left(P : K(p_0,p) \le \epsilon_n^2, V_+(p_0,p) \le \epsilon_n^2\right) \ge \exp(-n\epsilon_n^2 C), \tag{18}$$

where $N(\epsilon/2,\mathcal{P}_n,d)$ is the covering number, i.e., the minimum number of d-balls of size $\epsilon_n/2$ needed to cover $\mathcal{P}_n$.

We assume given values of J, K, and σ. We also make the simplifying assumption that the functional observations have already been detrended, so that $\mu(\cdot) \equiv 0$ and the prior mean, θ_0 is also taken to be 0. The theorem is stated in terms of the Frobenius norm on matrices. Now we state the main theorem of the paper on the rate of convergence of our posterior.

Theorem 1. *Let* $\boldsymbol{Y}_i \overset{iid}{\sim} N_T(\boldsymbol{0},\sigma^2\boldsymbol{I}+\boldsymbol{H}\boldsymbol{U}\boldsymbol{\Sigma}_0\boldsymbol{U}'\boldsymbol{H}')$, $i = 1,2,\ldots$ *for a known* σ^2 *and* $K = K_0 < T$ *fixed. Using the inverse Wishart prior from above*

$$\Pi\Big(\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}\|_F \ge M_n n^{-1/2}\log n|\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n\Big) \to 0,$$

as well as

$$\Pi\left(\boldsymbol{\Sigma} : \left\|\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}^{-1}\right\|_F \geq M_n n^{-1/2} \log n | \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\right) \to 0.$$

4.1 Proofs

Lemma 1 (Kullback–Leibler divergence under information loss)**.** *Let X and Y be random variables with densities p and q, respectively, and U be a random variable with uniform distribution on the unit interval, independent of X and Y. Let $\tilde{p}$ and $\tilde{q}$ be the densities of $T(X,U)$ and $T(Y,U)$, respectively, for a measurable function T. Then $K(\tilde{p},\tilde{q}) \leq K(p,q)$ and $V_+(\tilde{p},\tilde{q}) \leq V_+(p,q)$.*

This lemma can be proved by considering the conditional distributions of X given $T(X,U)$ and Y given $T(Y,U)$, and using the convexity of the maps $(u,v) \mapsto u \log \frac{u}{v}$ and $(u,v) \mapsto u \log^2 \frac{u}{v}$ on the set $u > v > 0$. A complete proof may be found in Ghosal and van der Vaart (2016, Appendix, Lemma B.12).

Next, we need a lemma relating the Hellinger distance to the Frobenius norm induced on our parameter of interest. This next lemma will be used in the entropy calculation.

Lemma 2. *Let d be the metric induced by the Hellinger distance on the k-dimensional centered multivariate Gaussian family. Then*

$$d(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = d_H(N(\mathbf{0}, \boldsymbol{\Sigma}_1), N(\mathbf{0}, \boldsymbol{\Sigma}_2)) \leq \left\|\boldsymbol{\Sigma}_1^{-1/2}\left(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1\right)\boldsymbol{\Sigma}_1^{-1/2}\right\|_F. \tag{19}$$

Furthermore, there exists $\delta > 0$ and constant $C > 0$, depending on $\boldsymbol{\Sigma}_1$, such that, if $d(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) < \delta$, then

$$\left\|\boldsymbol{\Sigma}_1^{-1/2}\left(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1\right)\boldsymbol{\Sigma}_1^{-1/2}\right\|_F \leq C d(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2). \tag{20}$$

Proof. Let $\{\lambda_j\}$ be the eigenvalues of $\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2}$, which is a symmetric, positive-definite matrix. Then for some orthogonal matrix $\boldsymbol{P}$. Then, $\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$. Then

$$\begin{aligned} d_H(N(\mathbf{0}, \boldsymbol{\Sigma}_1), N(\mathbf{0}, \boldsymbol{\Sigma}_2))^2 &= 1 - \frac{\det(\boldsymbol{\Sigma}_1)^{1/4}\det(\boldsymbol{\Sigma}_2)^{1/4}}{\det(\frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2))^{1/2}} \\ &= 1 - \frac{\det(\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2})^{1/4}}{\det(\frac{1}{2}(\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2} + \boldsymbol{I}))^{1/2}} \\ &= 1 - \frac{\prod_{j=1}^k \lambda_j^{1/4}}{\left(\prod_{j=1}^k \frac{1}{2}(\lambda_j + 1)\right)^{1/2}}. \end{aligned} \tag{21}$$

We now show that (21) is less than or equal to $\sum_{j=1}^k (\lambda_j - 1)^2$ by induction.

For $k = 1$, define a function $\xi(\lambda) = (\lambda - 1)^2 + \lambda^{1/4}(\frac{1}{2}(1+\lambda))^{-1/2}$. We have that $\xi(0) = \xi(1) = 1$, and we claim that $\xi(\lambda) \geq 1$ for $\lambda > 0$. Now,

$$\frac{d}{d\lambda}\xi(\lambda) = \frac{(\lambda - 1)(8\lambda^{3/4}(\lambda+1)^{3/2} - \sqrt{2})}{\lambda^{1/4}(\lambda+1)^{1/2}}. \tag{22}$$

The two positive real roots of the numerator are 1 and 0.403. It can then be seen that $\frac{d}{d\lambda}\xi(\lambda) \geq 0$ for $\lambda \leq 0.403$ or $\lambda \geq 1$, and that $\frac{d}{d\lambda}\xi(\lambda) \leq 0$ in between these values. Taking this together, this implies that $\xi(\lambda) \geq 1$ for all $\lambda \geq 0$, which is equivalent to the induction hypothesis for $k = 1$.

For the induction step, we define $\gamma(\lambda) = \lambda^{1/4}(\frac{1}{2}(1+\lambda))^{-1/2}$ and claim that $\gamma(\lambda) \leq 1$ for $\lambda \geq 0$. This can be seen by the fact that $\gamma(1) = 1$ and

$$\frac{d}{d\lambda}\gamma(\lambda) = -(\lambda - 1)2^{-3/2}\lambda^{-3/4}(\lambda+1)^{-3/2}$$

satisfies $\frac{d}{d\lambda}\gamma(\lambda) \geq 0$ for $0 \leq \lambda \leq 1$ and $\frac{d}{d\lambda}\gamma(\lambda) \leq 0$ for $\lambda \geq 1$. So, by the induction hypothesis and the case $k = 1$, we have that

$$\begin{aligned} 1 - \frac{\prod_{j=1}^{k}\lambda_j^{1/4}}{\left(\prod_{j=1}^{k}\frac{1}{2}(\lambda_j+1)\right)^{1/2}} &\leq 1 - \frac{\lambda_k^{1/4}}{(\frac{1}{2}(1+\lambda_k))^{1/2}} + \frac{\lambda_k^{1/4}}{(\frac{1}{2}(1+\lambda_k))^{1/2}}\sum_{j=1}^{k-1}(\lambda_j - 1)^2 \\ &\leq \sum_{j=1}^{k}(\lambda_j - 1)^2. \end{aligned} \tag{23}$$

The proof is now complete by noting that $\|\boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)\boldsymbol{\Sigma}_1^{-1/2}\|_F = \mathrm{tr}(\boldsymbol{P}(\boldsymbol{\Lambda} - \boldsymbol{I})\boldsymbol{P}') = \sum_{j=1}^{k}(\lambda_j - 1)^2$.

The reverse inequality can be established following the arguments given in Banerjee and Ghosal (2015, Lemma A.1). □

Proof of Theorem 1. Using the earlier notation, $p_0 = N(\boldsymbol{0}, \sigma^2\boldsymbol{I} + \boldsymbol{H}\boldsymbol{U}\boldsymbol{\Sigma}_0\boldsymbol{U}'\boldsymbol{H}')$ and $p = N(\boldsymbol{0}, \sigma^2\boldsymbol{I} + \boldsymbol{H}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}'\boldsymbol{H}')$. Let $q_0 = N(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$, $q = N(\boldsymbol{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1/2}$. Hence, by Lemma 1, we obtain that

$$\begin{aligned} &\Pi\left(P_0 \log\frac{p_0}{p} \leq \epsilon_n^2, P_0(\log_+\frac{p_0}{p})^2 \leq \epsilon_n^2\right) \\ &\geq \Pi\left(Q_0 \log\frac{q_0}{q} \leq \epsilon_n^2, Q_0(\log_+\frac{q_0}{q})^2 \leq \epsilon_n^2\right) \\ &= \Pi\left(\mathrm{tr}(\boldsymbol{\Sigma}^*) - k - \log\det(\boldsymbol{\Sigma}^*) < 2\epsilon_n^2, (\mathrm{tr}(\boldsymbol{\Sigma}^*) - k)^2 \leq 4\epsilon_n^2\right) \\ &= \Pi\left(\sum_{j=1}^{k}(\lambda_j - 1 - \log\lambda_j) < 2\epsilon_n^2, \left(\sum_{j=1}^{k}(\lambda_j - 1)^2\right)^2 \leq 4\epsilon_n^2\right), \end{aligned} \tag{24}$$

where $\{\lambda_j\}_{j=1}^k$ are the eigenvalues of the matrix $\boldsymbol{\Sigma}^*$. Now, $\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(v, \boldsymbol{\Xi})$ implies $\boldsymbol{\Sigma}^* \sim \text{Wishart}(v, \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Xi}\boldsymbol{\Sigma}_0^{1/2})$. Now, for large enough n,

$$\begin{aligned}
&\Pi\left(\sum_{j=1}^{k}(\lambda_j - 1 - \log\lambda_j) < 2\epsilon_n^2, \left(\sum_{j=1}^{k}(\lambda_j - 1)^2\right)^2 \leq 4\epsilon_n^2\right)\\
&\geq \Pi\left(\sum_{j=1}^{k}(\lambda_j - 1)^2 < 2\epsilon_n, \lambda_j \geq 1, j = 1, \ldots, k\right)\\
&\geq \Pi\left(1 \leq \lambda_j \leq 1 + k^{-1/2}\sqrt{2}\epsilon_n^{1/2}\right)\\
&\geq C_k \epsilon_n^{k^2/4+k/4} \gtrsim \exp(-n\epsilon_n^2 C) \qquad (25)
\end{aligned}$$

for $\epsilon_n = n^{-1/2}\log n$, and some constants, C and C_k. The first inequality of (25) is obtained by applying Lemma 1 of Shen et al. (2013); see Lemma 9.23 of Ghosal and van der Vaart (2016) for a complete proof, including the case $\boldsymbol{\Xi} \neq \boldsymbol{I}$.

Now, let $\mathcal{P}_n = \{\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma}^{-1}\|_F \leq n^{1/8}\}$. Let d be the distance on the space of symmetric, positive definite matrices induced by the Hellinger metric on the multivariate Gaussian family. For $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{P}_n$, we have that

$$d(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \leq \left\|\boldsymbol{\Sigma}_1^{-1/2}\left(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1\right)\boldsymbol{\Sigma}_1^{-1/2}\right\|_F \leq \|\boldsymbol{\Sigma}_1^{-1}\|_F\|\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1\|_F \leq n^{1/8}\|\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1\|_F, \qquad (26)$$

which implies that

$$\log N(\epsilon_n/2, \mathcal{P}_n, d) \leq \log N(n^{-1/8}\epsilon_n/2, \mathcal{P}_n, \|\cdot\|_F) \leq k^2 \log\left(\frac{6n^{1/4}}{\epsilon_n}\right) \lesssim n\epsilon_n^2 \qquad (27)$$

for the chosen ϵ_n. The second inequality on the preceding line is due to the fact that, using the Frobenius norm, the space of positive definite $k \times k$ matrices can be viewed as a subset of $\mathbb{R}^{k^2}$ (see Pollard (1990, Section 4) for the entropy calculation in Euclidean space).

Finally,

$$\begin{aligned}
\Pi\left(\|\boldsymbol{\Sigma}^{-1}\|_F > n^{1/8}\right) \leq \Pi\left(\text{tr}(\boldsymbol{\Sigma}^{-1}) > n^{1/8}\right) &\leq \exp(-n^{1/8})\det(\boldsymbol{I} - 2\boldsymbol{\Xi})^{-v/2}\\
&\lesssim \exp\left(-n\epsilon_n^2(C+4)\right) \qquad (28)
\end{aligned}$$

for the chosen ϵ_n. The first inequality is true because of the relationship between the Frobenius norm and the trace of a positive definite matrix. The second inequality follows from Markov's inequality using the moment generating function of $\text{tr}(\boldsymbol{\Sigma}^{-1})$ evaluated at 1 (Muirhead, 2009).

This establishing the consistency result in terms of the Hellinger distance, specifically, as $n \to \infty$,

$$\Pi\left(\boldsymbol{\Sigma} : d\left(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0\right) \geq M_n n^{-1/2}\log n | \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\right) \to 0.$$

Now, the reverse inequality of Lemma 2 implies the desired result in terms of the Frobenius norm:

$$\Pi\left(\boldsymbol{\Sigma}:\left\|\boldsymbol{\Sigma}_0^{-1/2}\left(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0\right)\boldsymbol{\Sigma}_0^{-1/2}\right\|_F \geq M_n n^{-1/2}\log n|\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n\right)\to 0.$$

Finally, since $\boldsymbol{\Sigma}_0$ is a fixed matrix, the first stated result follows from the relation $\|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F$, for two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. To derive the second assertion, write $\boldsymbol{\Sigma}_0^{-1}-\boldsymbol{\Sigma}^{-1}$ as $\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_0)\boldsymbol{\Sigma}^{-1}$ and apply the norm inequality repeatedly. □

5 Simulation Study

To assess the finite-sample performance of the proposed method, we present the results of a simulation study comparing the approach advocated herewithin to a recent frequentist approach, FACE (Xiao et al., 2013), implemented using the `refund` package in `R`.

5.1 Description of FACE

The frequentist method to which we compare our method is the "Fast Covariance Estimation" (FACE) method of Xiao et al. (2013). It is a very common frequentist method for the analysis of functional data, and has been made popular by the `refund` package available in `R`. The FACE estimator, $\tilde{F}$, is simply a sandwich-smoothed sample covariance matrix, that is

$$\tilde{F}=S\hat{F}S,$$

where $\hat{F}$ is the sample covariance matrix and S is a symmetric "smoothing" matrix, which is constructed using penalized B-splines. The form allows fast computation of the estimator. Thus, the comparison to our Bayesian method can be seen as demonstrating the potential benefits of a more complex method when the available computation allows for a fully Bayesian method.

5.2 Results

Each data set consists of 20 noisy observations on an evenly spaced grid of 50 time points in the interval $[-1,1]$. The true underlying functional observations all have a true mean of $\mu(t)=\sin(2\pi t)$, and a covariance of either $\kappa_1(s,t)=\exp\{-3(t-s)^2\}$ or $\kappa_2(s,t)=\min\{s+1,t+1\}$, depending on the experimental conditions. Independent sampling noise is then added in the form of independent $N(0,0.3)$ random variables (other values for the variance of the noise were considered, but the results remained qualitatively very similar). The two methods are compared in the following realms: estimation of the mean function, estimation of the covariance function, estimation of the principal components, and reconstruction of a new set of underlying functional observations (generated according to the same model). Function estimation is evaluated using the supremum, L_1 and L_2 metrics, and principal component estimation is evaluated using the angle between the estimate and the true function (this was chosen instead of

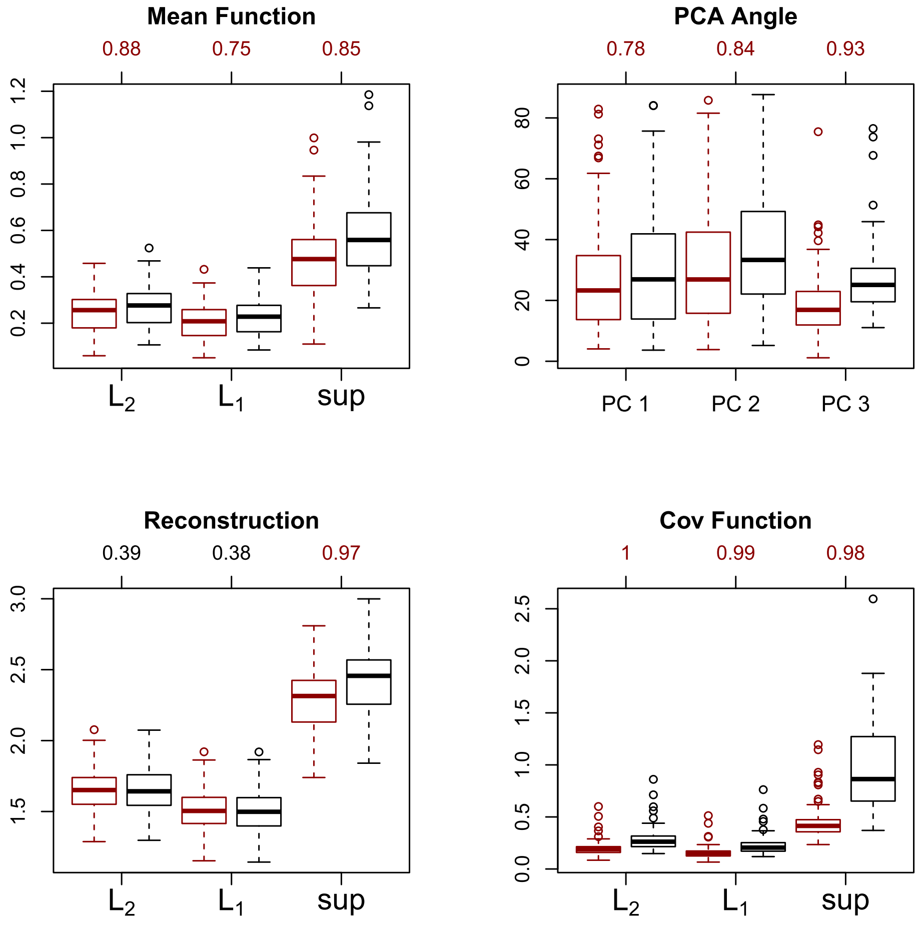

Figure 1: Simulation Study: Boxplots for the simulation results corresponding to κ_1 used for both the truth and the prior. Results are shown in pairs, with the left box (red color) representing the proposed Bayesian procedure, and the right (black) representing the FACE procedure. The value above each pair represents the proportion of data sets in which the Bayesian procedure performed superiorly.

using squared distance, to take advantage of the Hilbert space structure). Observations are always reconstructed using the first 4 principal components.

For each choice of κ_1 and κ_2 as the true covariance function, we ran two simulations corresponding to choosing κ_1 or κ_2 as the prior covariance, yielding four total experimental conditions. Each scenario is repeated 100 times.

The proposed Bayesian procedure is computed using a Gibbs sampler with 5000 burn-in iterations, and 5000 iterations to estimate posterior means. The concentration parameter for the Wishart distribution, ν, was chosen to be $2K$; this empirically seemed

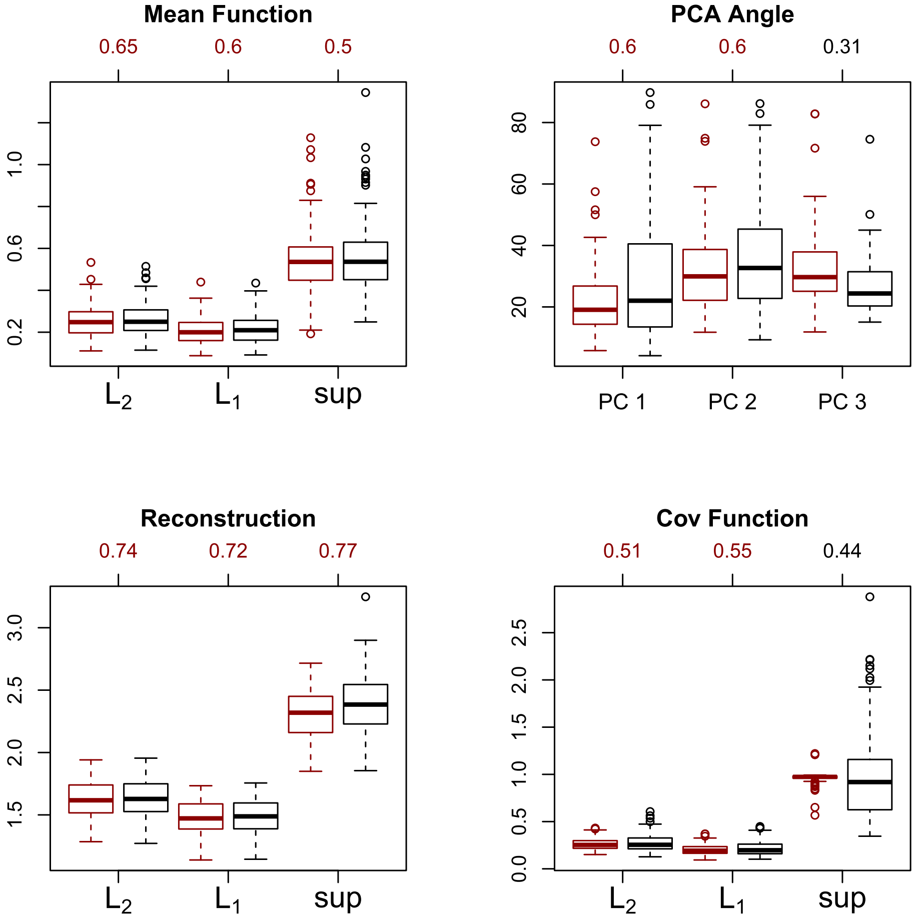

Figure 2: Simulation Study: Boxplots for the simulation results corresponding to κ_1 used for the truth and κ_2 used for the prior. Results are shown in pairs, with the left box (red color) representing the proposed Bayesian procedure, and the right (black) representing the FACE procedure. The value above each pair represents the proportion of data sets in which the Bayesian procedure performed superiorly.

to be a reasonable default choice. Legendre polynomials were used as the orthonormal basis, a Poisson prior with mean 7 was placed on the number of basis functions, and a Poisson prior with mean 1 was placed on the number of principal components. However, for the computation, only models with fewer than 20 basis functions were run, which would correspond to a truncated prior. This practically had no effect on the results since models outside that range had negligible posterior mass. See Figure 5 for examples of the posterior distribution of (J, K). The principal components were estimated using the decomposition of the estimated posterior mean of the covariance matrix. The results can be seen in Figures 1–4.

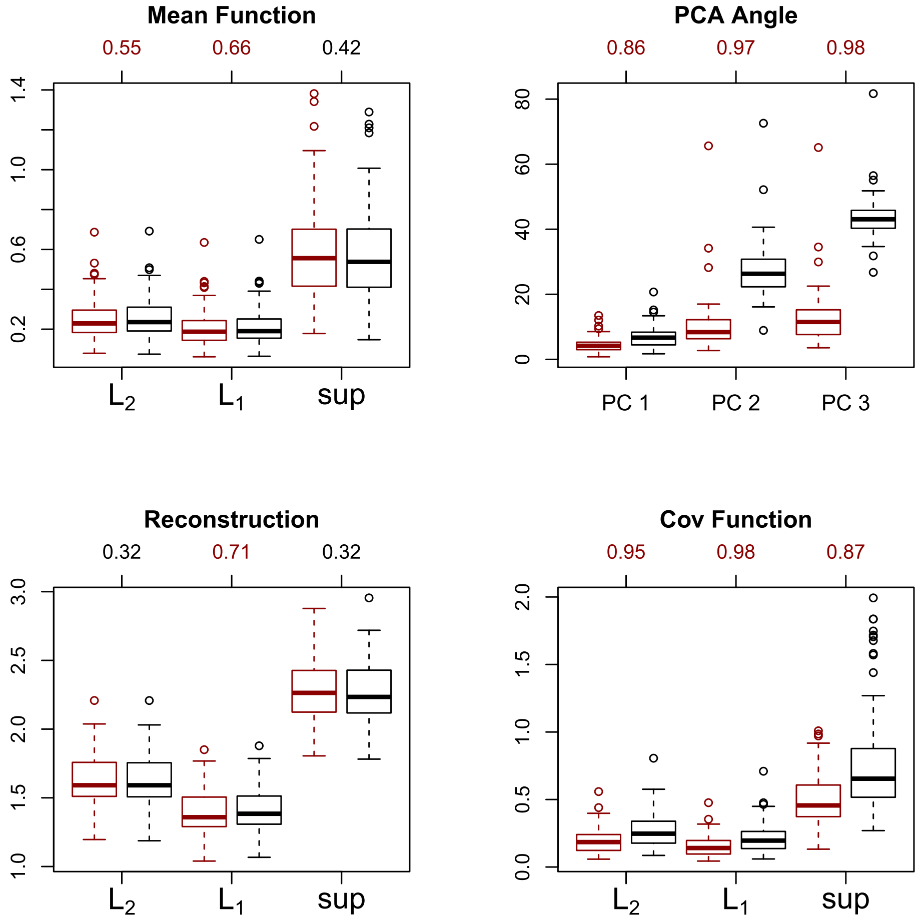

Figure 3: Simulation Study: Boxplots for the simulation results corresponding to κ_2 used for both the truth and the prior. Results are shown in pairs, with the left box (red color) representing the proposed Bayesian procedure, and the right (black) representing the FACE procedure. The value above each pair represents the proportion of data sets in which the Bayesian procedure performed superiorly.

As can be seen from the results, the Bayesian procedure performs consistently well across the conditions in the estimation of the principal components themselves when measured by the angle from the truth. The practical importance of prior information can be seen in the improvements in reconstruction when the true covariance is used to construct the prior. In the functional data setting, the smoothness of the underlying true observations is usually well understood scientifically in an applied context, and should be incorporated into the analysis. The overall picture that these results show is that the proposed Bayesian method has the ability to perform competitively with the most modern frequentist procedures when judged by repeated sampling criteria.

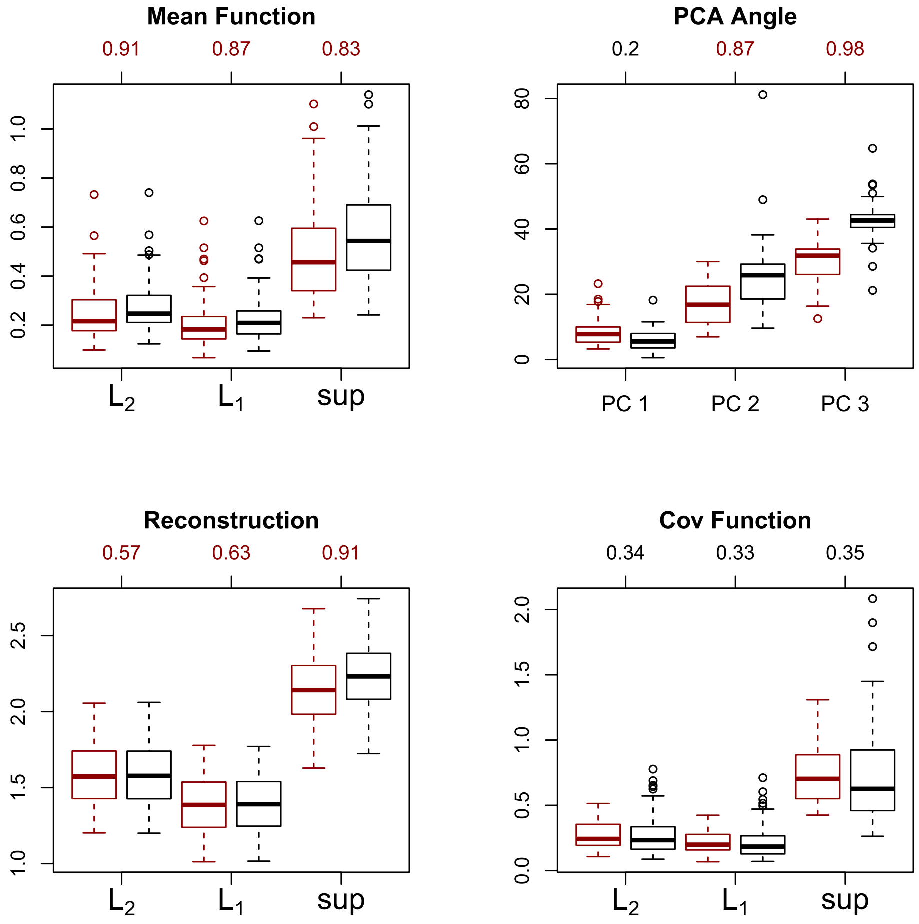

Figure 4: Simulation Study: Boxplots for the simulation results corresponding to κ_2 used for the truth and κ_1 used for the prior. Results are shown in pairs, with the left box (red color) representing the proposed Bayesian procedure, and the right (black) representing the FACE procedure. The value above each pair represents the proportion of data sets in which the Bayesian procedure performed superiorly.

6 Canadian Weather Data

To illustrate our method on real data, we analyzed the popular Canadian weather data, which is freely available in the `fda` package in `R`. The data was made popular by Ramsay and Silverman (2005), and our analysis is consistent with theirs. These data consist of 35 functional observations observed on a common grid of 365 time points. They correspond to the average daily temperature of 35 Canadian cities. We employ Legendre polynomials as the basis, with an unknown number of basis functions. We use a modified Poisson distribution on K with mean 7, and truncated above at 30. The

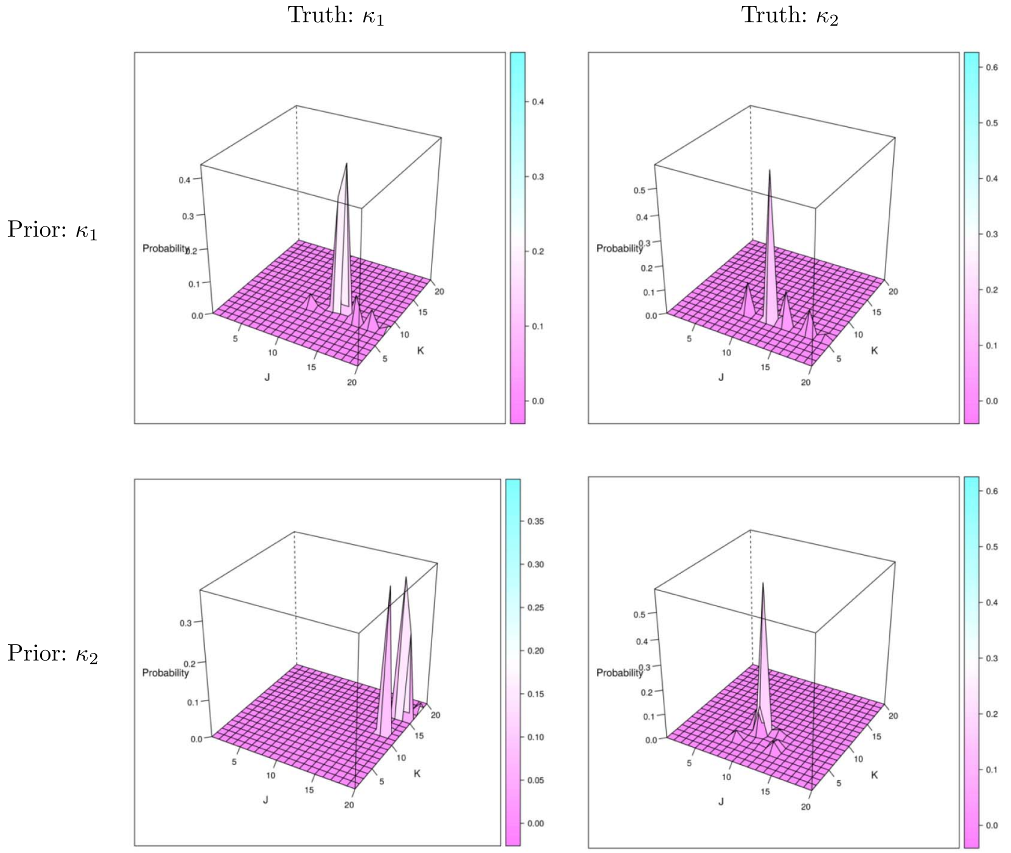

Figure 5: Simulation Study: Examples of the posterior distribution of models for each of the four experimental conditions. Row 1 corresponds to κ_1 as the prior, and column 1 corresponds to κ_1 as the truth. The other row and column correspond to κ_2.

parameter $\boldsymbol{\theta}_0$ is taken to be zero, and the prior covariance function that is approximated was $\kappa(s,t) = \exp\{-3(s-t)^2\}$. For each model, 180,000 MCMC iterations were used for estimation after 20,000 burn-in iterations. Estimates are only calculated at the sampled time points; if there are other time points of interest, ideally, they should be treated as missing data and incorporated into the MCMC approximations.

In the posterior, almost all the mass lies on the model with $J = 12, K = 12$ (96.1%), and a small amount on $J = 16, K = 15$ (3.9%); a plot of the marginal likelihoods for each model can be seen in Figure 6. Posterior estimates shown are full model estimates, although they will be extremely close to conditioning on the maximum *a posteriori* model. The observations along with their smoothed estimates can be seen in Figure 7. The estimated covariance function can be seen in Figure 8, along with the implied principal components in Figure 9.

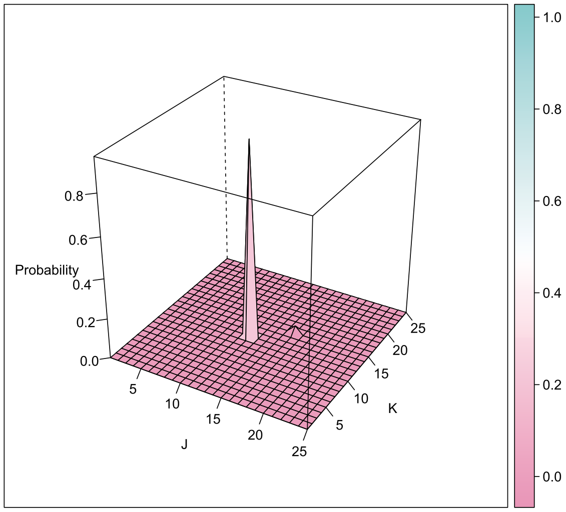

Figure 6: Canadian Weather Data: Posterior probabilities for each model, (J, K). Only models $1 \leq K \leq J \leq 25$ had prior mass.

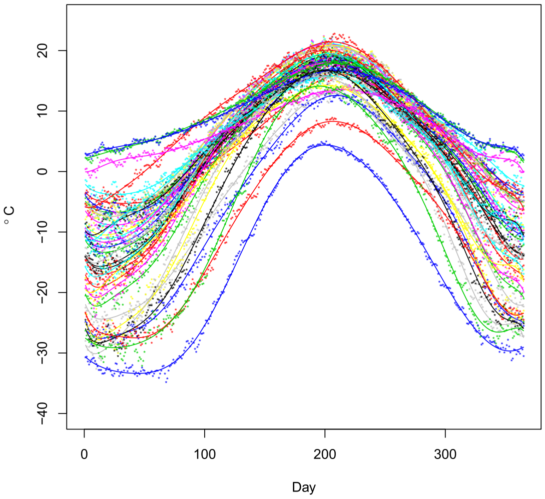

Figure 7: Canadian Weather Data: Observations (dots) along with pointwise posterior means (lines). The posterior means are the mean of $\boldsymbol{H}_J \boldsymbol{U}_K \boldsymbol{\beta}_{i,K}$ over (J, K).

The first principal component represents the overall temperature of the city throughout the year; it differentiates between generally "mild" and "cold" cities. The second principal component seems to quantify the relative difference in temperature between summer and winter months, and differentiates between cities that have a more flat temperature function, compared to those with extremely cold winters. The higher order principal components represent more complicated phenomena.

	Prior	Truth	MSE(Bayes)	MSE(FACE)	P(MSE(Bayes) < MSE(FACE))
PC1	κ_1	κ_1	0.066	0.075	0.78
		κ_2	0.020	0.016	0.26
	κ_2	κ_1	0.053	0.068	0.60
		κ_2	0.011	0.017	0.89
PC2	κ_1	κ_1	0.073	0.088	0.84
		κ_2	0.048	0.069	0.92
	κ_2	κ_1	0.076	0.086	0.60
		κ_2	0.025	0.071	0.98
PC3	κ_1	κ_1	0.045	0.065	0.93
		κ_2	0.074	0.101	0.98
	κ_2	κ_1	0.078	0.064	0.31
		κ_2	0.031	0.104	0.98

Table 1: Simulation Study: Comparison of MSEs (Mean Squared Errors). Rows in dark red text are experimental conditions where our method outperformed FACE. In all of these cases the difference is significant in a frequentist sense with the null hypothesis that there is equal probability of either method winning a trial.

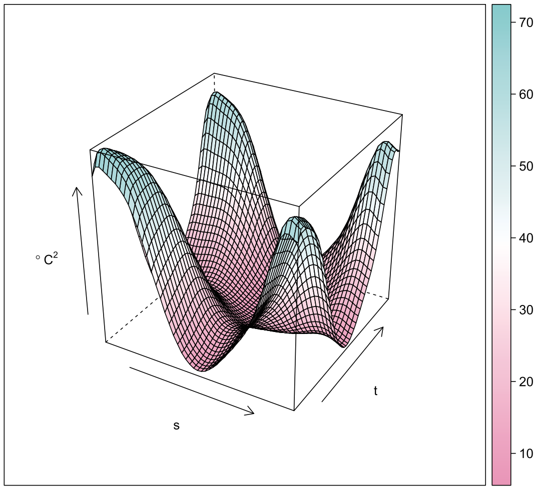

Figure 8: Canadian Weather Data: Full posterior mean of the covariance function over all (J, K) pairs.

7 Discussion

There are two points that warrant further discussion beyond what has already been presented: modifications to the model under special circumstances, and the computational difficulties of fitting the model.

Although our method allows for the possibility that the number of basis functions used for approximation and the number of principal components require different values,

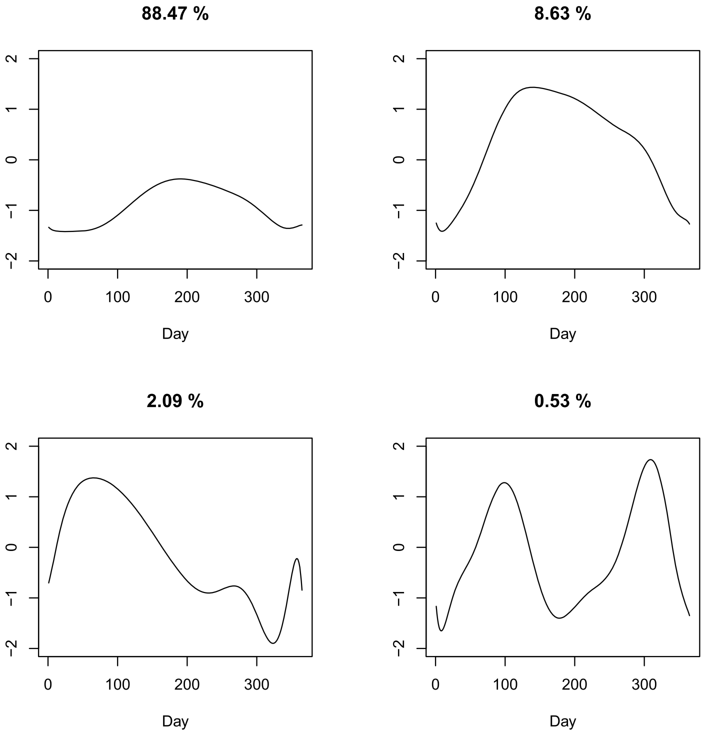

Figure 9: Canadian Weather Data: First four FPCAs corresponding to the full posterior mean of the covariance function (Figure 8) along with their associated proportion of variance explained.

in the form we present, the number of basis functions is the same for both mean and covariance modeling. One possible concern is that, in the situation where the mean function requires many more basis functions to be well-approximated compared with the covariance function, the chosen value of K will be forced too high. Specifically, to deal with this, we can allow for an extra overall mean term, $\boldsymbol{\eta}$, in (10) that can capture the excess roughness present in the mean:

$$\boldsymbol{Y}_i \overset{ind}{\sim} N_T(\boldsymbol{H}_J \boldsymbol{U}_K \boldsymbol{\beta}_{i,K} + \boldsymbol{\eta}, \sigma^2 \boldsymbol{I}) \tag{29a}$$

$$\boldsymbol{\eta} \sim N_T(\boldsymbol{\eta}_0, \boldsymbol{\Psi}). \tag{29b}$$

Allowing $\boldsymbol{\Psi}$ to be full rank would be sure to provide great flexibility, but may lead to over-fitting. A further modification could be to express $\boldsymbol{\eta}$ in the same basis domain as the rest of the model, and let $\boldsymbol{\eta} = \boldsymbol{H}_M \boldsymbol{\xi}$, $M > K$, with a prior on $\boldsymbol{\xi}$. This is similar to the approach in generalized additive models of allowing differing number of basis functions for each component, and the addition of this term causes an identifiability issue with the collection $\{\boldsymbol{\beta}_{i,K}\}$. This issue can be handled in a similar fashion to Lang and Brezger (2004). A step-wise frequentist approach could be to de-mean the data before processing, possibly including smoothing, and then apply our method with θ removed from the model, that is, $\boldsymbol{\beta}_{i,K} \overset{iid}{\sim} N_K(\mathbf{0}, \boldsymbol{\Sigma})$ in (10).

The final issue is the computational difficulties in fitting the full model. The approach we have taken is to run independent MCMC chains for each pair of values (J, K), $1 \leq K \leq J \leq J_{\max}$, up to some predefined value, $J_{\max}$. The software approach the we have implemented will be made available on one of the authors' website (`https://www.ajsuarez.com`). It uses `R`'s `C` interface to be able to take advantage of using OpenMP for parallelization of the MCMC chains. Specifically, for each value of J, the models, $1 \leq K \leq J$, are run in parallel batches. For example of computation time, on a data set containing 50 functional observations at 100 time points, fitting all models up to $J_{\max} = 25$ takes approximately 3 seconds per 1000 MCMC steps on a 6-core Intel Haswell CPU running at 4 GHz. Computation increases on the order of $J^2_{\max}$, and this dominates compared with the time points. For an MCMC procedure that provides full information on each individual model, we believe this approach to be worth the computational time.

References

Ash, R. (1965). *Information Theory*. Interscience Tracts in Pure and Applied Mathematics. Interscience Publishers. MR0229475. 312, 313

Banerjee, S. and Ghosal, S. (2015). "Bayesian structure learning in graphical models." *Journal of Multivariate Analysis*, 136: 147–162. MR3321485. doi: `http://dx.doi.org/10.1016/j.jmva.2015.01.015`. 321

Behseta, S., Kass, R. E., and Wallstrom, G. L. (2005). "Hierarchical models for assessing variability among functions." *Biometrika*, 92(2): 419–434. MR2201368. doi: `http://dx.doi.org/10.1093/biomet/92.2.419`. 311

Bishop, C. M. (1999). "Bayesian PCA." *Advances in Neural Information Processing Systems*, 382–388. 311

Chib, S. (1995). "Marginal Likelihood from the Gibbs Output." *Journal of the American Statistical Association*, 90(432): 1313–1321. MR1379473. 316

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). "Convergence rates of posterior distributions." *Annals of Statistics*, 28(2): 500–531. MR1790007. doi: `http://dx.doi.org/10.1214/aos/1016218228`. 319

Ghosal, S. and van der Vaart, A. (2016). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, to appear. 320, 322

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. MR1380810. doi: `http://dx.doi.org/10.1093/biomet/82.4.711`. 316

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, volume 200. Springer. MR2920735. doi: `http://dx.doi.org/10.1007/978-1-4614-3655-3`. 311

Jolliffe, I. T. (1986). *Principal Component Analysis*, volume 487. Springer-Verlag New York. MR2036084. 311

Lad, F. (1996). *Operational Subjective Statistical Methods: a Mathematical, Philosophical, and Historical Introduction*. Wiley Series in Probability and Statistics. Wiley. MR1421323. 318

Lang, S. and Brezger, A. (2004). "Bayesian P-splines." *Journal of Computational and Graphical Statistics*, 13(1): 183–212. MR2044877. doi: http://dx.doi.org/10.1198/1061860043010. 331

Muirhead, R. J. (2009). *Aspects of Multivariate Statistical Theory*. Wiley–Interscience. MR0652932. 322

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. NSF–CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics. MR1089429. 322

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer. MR2168993. 311, 327

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures." *Biometrika*, 100(3): 623–640. MR3094441. doi: http://dx.doi.org/10.1093/biomet/ast015. 322

Tipping, M. E. and Bishop, C. M. (1999). "Probabilistic principal component analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3): 611–622. MR1707864. doi: http://dx.doi.org/10.1111/1467-9868.00196. 311

van der Linde, A. (2008). "Variational Bayesian functional PCA." *Computational Statistics & Data Analysis*, 53(2): 517–533. MR2649106. doi: http://dx.doi.org/10.1016/j.csda.2008.09.015. 312

Xiao, L., Ruppert, D., Zipunnikov, V., and Crainiceanu, C. (2013). "Fast covariance estimation for high-dimensional functional data." arXiv:1306.5718. MR3439382. doi: http://dx.doi.org/10.1007/s11222-014-9485-x. 323

Acknowledgments

Research is partially supported by NSF grant number DMS-1106570.