Integrated Model Selection and Scalability in Functional Data Analysis through Bayesian Learning

Wenzheng Tao , Sarang Joshi * , Ross Whitaker

Posted Date: 10 March 2025

doi: 10.20944/preprints202503.0658.v1

Keywords: Functional data analysis; Principal component analysis; Dimension reduction; Sparse Bayesian learning; Variational Bayesian inference; Nonparametric methods; Model selection

 doi:10.20944/preprints202503.0658.v1

Article

Integrated Model Selection and Scalability in Functional Data Analysis through Bayesian Learning

Wenzheng Tao [1,2], Sarang Joshi [2,3,*] and Ross Whitaker [1,2]

[1] School of Computing, The University of Utah;
[2] Scientific Computing and Imaging Institute, The University of Utah;
[3] Biomedical Engineering, The University of Utah;
* Correspondence: sarang.joshi@utah.edu

Abstract: Functional data, including one-dimensional curves and higher-dimensional surfaces, have become increasingly prominent across scientific disciplines. They offer a continuous perspective that captures subtle dynamics and richer structures compared to discrete representations, thereby preserving essential information and facilitating more natural modeling of real-world phenomena, especially in sparse or irregularly sampled settings. A key challenge lies in identifying low-dimensional representations and estimating covariance structures that capture population statistics effectively. We propose a novel Bayesian framework with a nonparametric kernel expansion and a sparse prior, enabling direct modeling of measured data and avoiding the artificial biases from regridding. Our method, Bayesian scalable functional data analysis (BSFDA), automatically selects both subspace dimensionalities and basis functions, reducing computational overhead through an efficient variational optimization strategy. We further propose a faster approximate variant that maintains comparable accuracy but accelerates computations significantly on large-scale datasets. Extensive simulation studies demonstrate that our framework outperforms conventional techniques in covariance estimation and dimensionality selection, showing resilience to high dimensionality and irregular sampling. The proposed methodology proves effective for multidimensional functional data and showcases practical applicability in biomedical and meteorological datasets. Overall, BSFDA offers an adaptive, continuous, and scalable solution for modern functional data analysis across diverse scientific domains.

Keywords: functional data analysis; principal component analysis; dimension reduction; sparse bayesian learning; variational bayesian inference; nonparametric methods; model selection

1. Introduction

The emergence of big data across diverse fields, such as biomedicine, finance, and physical modeling, has catalyzed the need for advanced analytical methodologies capable of handling complex, high-dimensional datasets that conventional discrete-data analysis approaches cannot always process effectively. Such datasets often require analysis that captures and interprets their continuous and potentially high-dimensional complexities–a central promise of *functional data analysis* (FDA) [1,2]. Foundational work established FDA's capacity to treat each observation as an entire function [3], be it a curve, surface, or higher-dimensional structure, thereby extracting richer insights than conventional discrete-point analyses. Over the past decade, FDA's scope has widened significantly to accommodate high-dimensional and multivariate applications with theoretical and computational advances emerging across various contexts [4–6].

A pivotal technique within FDA is *functional principal component analysis* (fPCA), which serves as a dimension-reduction tool similar to classical PCA and factor analysis. Unlike classical PCA, however, fPCA operates in principle in an infinite-dimensional function space to capture dominant modes of variation and reduce complexity [7]. Despite its conceptual elegance, existing fPCA and similar FDA models often assume that data is observed on a shared, finite grid, often relying on heuristic imputation or posterior estimation to handle any missing entries [8–18]. This assumption conveniently facilitates the adoption of established linear algebraic methods, but compromises the integrity of FDA

 doi:10.20944/preprints202503.0658.v1

by introducing significant information loss and high computational demands in high-dimensional applications.

Ideally, each function would be represented according to its naturally sampled measurement points rather than forcing all observations onto a shared grid, thus preserving crucial information and avoiding the need for heuristic resampling. This point is critical when considering that, given only a finite number of data points, infinitely many functions can interpolate these points, each reflecting different inductive biases about smoothness or shape [19]. Conventional smoothing or regridding methods (e.g., polynomial interpolation) introduce biases that may distort the underlying function's actual behavior. In contrast, we will achieve more accurate and unbiased predictions by concurrently updating the function estimation and the population-level statistics governing the estimation, such as those encoded in the covariance operator. Such an approach requires directly modeling the function from its original measurement points rather than imposing artificial grids.

To mitigate these limitations, several studies have proposed alternative strategies. For instance, [3] developed a nonparametric technique for estimating mean and covariance for functional data under smoothness assumptions while also discussing a continuous formulation and the necessary discretization in practical applications. In [20,21], they extended fPCA to sparse and irregular longitudinal designs by smoothing the covariance estimate and then discretizing. Nonetheless, classical discretization steps often result in significant information loss and computational burdens.

As functional data size and complexity grew, researchers turned to flexible basis expansions, including sinusoids (Fourier), wavelets, polynomials, and B-splines, for a finite-dimensional representation of functional data that is convenient and accurate in computation, avoiding the drawbacks of explicit approximation and resampling [2,22–26]. For example, [27] utilized basis function approximations to manage irregular grids. However, a core challenge remains in selecting a suitable model. For instance, researchers must choose the number and form (e.g., smoothness), along with the dimensionality of the representational subspace. In approximation, the placement of basis functions is also essential. Evenly spaced nodes remain popular for their simplicity but may be suboptimal. Alternative node allocations may be better, like Chebyshev nodes for superior accuracy [28], or sparse grids to reduce combinatorial growth of computational complexities [29].

Existing studies tend to rely on choosing the hyperparameters manually [7,10,11,30], or cross-validation [3,25,31,32], which are known to be computationally prohibitive. Others employ approximated cross-validation [22,24] or marginal likelihood [14] but still require exhaustive testing of all candidate models. Methods with sparse Bayesian priors [8,33,34] for model selection allow model selection with a single optimization. In [35,36], they use shrinkage or sparse priors for data-adaptive basis selection to ensure minimal but effective sets of basis functions. Notably, [37] proposed Bayesian and Akaike information criterion demonstrating state-of-the-art performance in simulation studies for sparse and dense functional data.

In addition, probabilistic FDA emerges as a sophisticated adaptation of probabilistic methods tailored to incorporate the flexibility of latent variable models to manage functional data. A Bayesian latent factor regression model (LFRM) [34], for example, extends conventional regression to accommodate complex structures and dependencies in functional data, providing a robust framework to handle the complexities inherent in functional data. However, these Bayesian approaches are often limited with the computational demands of Monte Carlo methods in high dimensions [14]. To address increasingly high-dimensional FDA problems, recent efforts have emphasized scalability. For instance, [7,30,38,39] introduced FDA for 2D and 3D images with a fixed basis or grid. In [32], they further reduced complexity in 2D fPCA via tensor product B-splines. And [40] applied a Bayesian framework with basis expansion, adaptive regularization, and Gibbs sampling to 2D functional data in the form of EEG studies on children with autism. Further, [41] leverages a parsimonious basis representation and variational Bayes to achieve computational efficiency, making it suitable for 3-D brain imaging data.

In parallel, the broader field of principal component analysis (PCA) remains a fundamental and effective tool. Classical PCA, rooted in eigen-decomposition [42], effectively extracts dominant

 doi:10.20944/preprints202503.0658.v1

modes of variation in many settings but does not inherently accommodate the probabilistic nature of real-world data and its inherent uncertainties. Thus, [43] introduced *probabilistic PCA* (PPCA), which incorporates a probability distribution to manage these uncertainties more effectively. PCA has since evolved to address missing data, model selection, and complex data types [44–47]. In the context of functional data, these concepts motivate new approaches that unify probabilistic methodologies, latent factor models, and kernel expansions for continuous domains [1].

Within Bayesian machine learning, various priors have been proposed for sparse or robust formulations of PCA [48–51]. Specifically, *sparse Bayesian learning* (SBL) [52–54] with its mechanism automatic relevance determination (ARD) [55,56] has proven adept at promoting parsimonious solutions [57–61]. SBL has emerged in Bayesian PCA [45,62,63], applying an iterative method to evaluate the relevance of each component and select the internal dimensionality by disregarding the redundant ones. [64] applied SBL to optimize the combination of base kernels to enhance model performance. These methods often exploit variational techniques or accelerated optimization [17,64–70], thereby balancing model complexity with computational tractability. In functional data contexts, where representations are infinite-dimensional, SBL offers a compelling framework for advanced FDA methods by efficiently handling sparse expansions and adaptively adjusting model complexity.

In summary, despite these efforts to advance functional data analysis, several challenges persist. Existing methods often exhibit limitations in accuracy and efficiency when sampling is sparse, automatic model selection is essential, and dimensionality is high [41]. Concurrently, probabilistic PCA and SBL frameworks illustrate powerful strategies to incorporate versatility and adaptivity for such data complexities while their adaptation to FDA is still evolving. These gaps underscore a need for a robust, flexible, and computationally feasible approach, unifying ideas from FDA, PPCA, and SBL, that manages the continuous and high-dimensional intricacy of modern datasets.

1.1. Contriubutions

This manuscript proposes a novel *Bayesian framework for functional principal component analysis* that leverages nonparametric kernel expansions, sparse Bayesian learning for model selection, and efficient variational inference (VI). We abbreviate the proposed method as *BSFDA* (Bayesian Scalable Functional Data Analysis). *BSFDA* addresses critical gaps in existing FDA techniques with irregular sampling, high-dimensional scalability, and selection of both basis functions and principal components. Specifically, our approach offers:

- **Joint selection of optimum latent factors and sparse basis functions**: This eliminates constraints on parametric representation dimensionality, avoids information loss from discretization, and extends naturally to higher dimensions or non-Euclidean spaces through nonparametric kernel expansion. It further enhances interpretability by adaptively choosing model complexity without testing multiple models separately. We achieve these improvements using a Bayesian paradigm that provides robust and accurate posterior estimates while supporting uncertainty quantification [1].
- **Scalability across domain dimensionality and data size**: The proposed method uses VI for faster computation compared to Markov chain Monte Carlo (MCMC) methods, while still being accurate. BSFDA reduces overall computation by partitioning the parameters into smaller update groups, and introducing a slack variable to further subdivide the weighting matrix (which is part of the kernel structure) into even smaller parts [34], updating fewer blocks at a time and considering all model options. Introducing a slack variable makes the optimization process more efficient by separating different variable groups. This approach scales well with data size and works efficiently even with large, complex datasets. We demonstrate this on the 4D global oceanic temperature data set (ARGO), which consists of 127 million data points spanning across the globe for 27 years with depths up to 200 meters [71].

 doi:10.20944/preprints202503.0658.v1

1.2. Outline

Together, these contributions position our work at the intersection of functional principal component analysis [25] and sparse Bayesian learning [52], enabling robust, flexible, and computationally feasible analysis of high-dimensional functional data. The remainder of this paper is organized as follows. We first describe the proposed Bayesian functional PCA framework in detail, highlighting the nonparametric kernel expansions and sparse Bayesian priors. Next, we discuss the variational inference procedure and the reduced active block updating step, illustrating how these techniques jointly provide scalability and accuracy. We then present extensive empirical studies demonstrating factor selection accuracy, covariance operator estimation, and performance in large-scale 4D applications. Finally, we conclude with a discussion of potential extensions and open directions, emphasizing the broader implications of our work for large-scale, high-dimensional functional data analysis.

2. Formulation

The aim of the proposed method is the estimation of functions $y_i : R^M \mapsto R$ that are outcomes of an M-dimensional stochastic process. The observed data are P independent, noisy samples of the functions $\{y_i\}_{i=1}^P$ at index $\{X_i \in R^{N_i \times M}\}_{i=1}^P$ where N_i is the number of measured samples for the i^{th} function y_i and $X_{in} \in R^M$ is the location of the n^{th} measurement in the domain of the sample. The observations are $\{Y_i\}_{i=1}^P$, where $Y_{in} = y_i(X_{in}) + E_{in}$, where E_{in} is white Gaussian noise of variance σ^2.

2.1. Generative model

We assume that y_i is in a class of functions that can be approximated as a weighted summation of K kernel functions $\{\phi_k\}_{k=1}^K$:

$$y_i(x) = \sum_{k=1}^{K} w_{ik}\phi_k(x), \tag{1}$$

where $\phi_k(x) = \mathcal{K}(x, X_{in})$, $\mathcal{K}$ is the kernel function, X_{in} is the k-th location, $\{w_{ik}\}_{k=1}^K$ are the K coefficients of y_i. We also assume that the functions span a low dimensional subspace of dimension $J << K$. We model this stochastically by assuming that the weights, $w_i \in R^K$, are given by $w_{ik} = \sum_{j=1}^J Z_{ij}W_{jk} + \bar{Z}_k$, where $W \in R^{J\times K}$ are the principal component loadings and $Z_i \in R^J$ are standard normal variables. This model is therefore:

$$Y_{in} = \sum_{k=1}^{K}\left(\left(\sum_{j=1}^{J}\left(Z_{ij}W_{jk}\right) + \bar{Z}_k\right)\phi_k(X_{in})\right) + E_{in} = (Z_iW + \bar{Z})\Phi_{i\cdot n} + E_{in}, \tag{2}$$

where $\Phi_{i\cdot n} = [\phi_1(X_{in}), \ldots, \phi_K(X_{in})]^T$ are the evaluations of the basis functions at the n-th index of the i-th sample function.

2.2. Sparse Prior

For effective model selection, we introduce a sparse prior over the coefficients of the basis functions. The sparse prior in the proposed model is based on automatic relevance detection (ARD) [62]. ARD evaluates the importance of a feature with a precision parameter estimated from the data. The model uses $\{\alpha_j\}_{j=1}^J$ and $\{\beta_k\}_{k=1}^K$ for the numbers of components and basis functions, respectively, while η signifies the overall magnitude of the mean coefficients:

$$\bar{Z}_k \sim \mathcal{N}(0, \eta^{-1}\beta_k^{-1}), \forall k = 1 : K \tag{3}$$

$$W_{jk} \sim \mathcal{N}(0, \alpha_j^{-1}\beta_k^{-1}), \forall j = 1 : J, k = 1 : K \tag{4}$$

 doi:10.20944/preprints202503.0658.v1

In the model, $\alpha_j, \beta_k, \eta, \sigma^{-2}$ are all variables of precision parameters, coming naturally with a conjugate prior of Gamma distribution. The probabilistic graphical model is depicted in Figure 1. Setting a_0, b_0 to a small value yields a vague Gamma prior that approximates a noninformative (Jeffreys-type) prior.

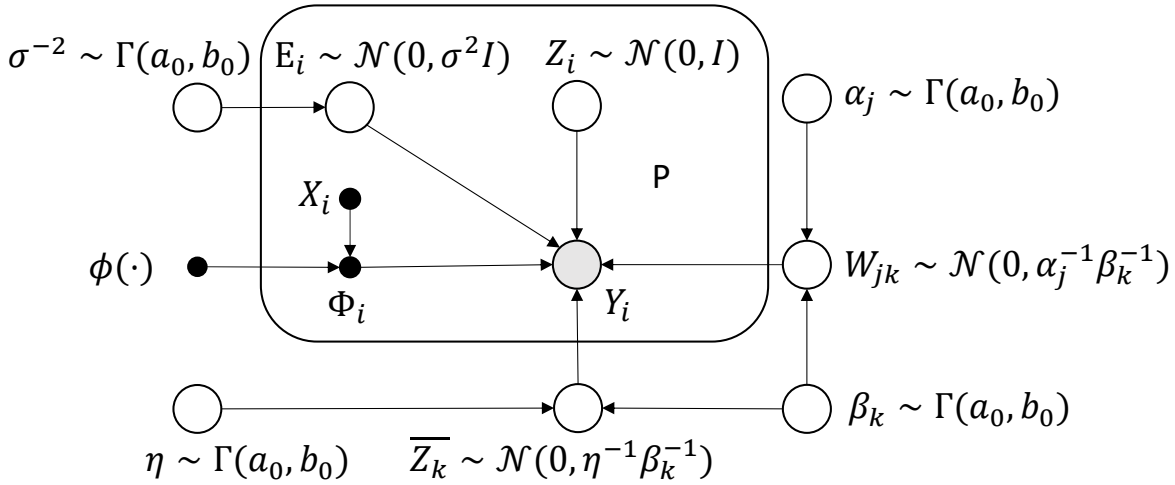


Figure 1. Probabilistic graphical model for the full model.

3. Methods

Based on the proposed formulation in Section 1.2, we estimate $\Pr[\Theta|X, Y, a_0, b_0]$, the posterior of the unobserved values $\Theta = \{Z, W, \bar{Z}, \sigma, \alpha, \beta, \eta\}$. This inference gives the point estimates of Θ and the posterior predictive distribution of new data. For notational convenience, X, a_0, b_0 are omitted.

Using Bayes' theorem, $\Pr[\Theta|Y] = \frac{\Pr[Y|\Theta]\Pr[\Theta]}{\Pr[Y]}$, but the exact posterior distribution is intractable because the evidence $\Pr[Y] = \int \Pr[\Theta, Y] d\Theta$ is intractable. Therefore, an approximate inference strategy is proposed. To facilitate this, we utilize variational inference (VI) [72], choosing a surrogate density from a parameterized family, denoted as $\mathcal{Q}$, to approximate the posterior. Compared with classical methods like Markov chain Monte Carlo (MCMC) sampling, VI is typically faster per [72]. In our experiments, VI is about 85 times faster for the original Bayesian PCA formulation [45] as shown in Section 12.2 in the supplements.

3.1. Variational Bayesian Inference

Variational inference optimizes $\mathcal{Q}$ by maximizing the lower bound $\mathcal{L}$ (minimizing the KL divergence between actual and surrogate distributions):

$$\mathbb{E}_{\mathcal{Q}}\left[\ln \frac{\Pr[\Theta, Y]}{\mathcal{Q}(\Theta)}\right] = -\mathrm{KL}(\mathcal{Q}(\Theta)||\Pr[\Theta|Y]) + \ln \Pr[Y] \propto -\mathrm{KL}(\mathcal{Q}(\Theta)||\Pr[\Theta|Y]). \tag{5}$$

The mean-field variational family is used for $\mathcal{Q}$. It simplifies the optimization by assuming the surrogate posterior distributions are independent, allowing each variable in the posterior to be optimized independently: $\mathcal{Q}_\Theta = \prod_i \mathcal{Q}_{\Theta_i}$. The posterior for each variable is chosen conjugate, further simplifying the optimization. Thus, the posteriors of the component scores Z, the weighting matrix W, and the mean weights $\bar{Z}$ are normal distributions. Here W is vectorized via vec (W) without altering its

 doi:10.20944/preprints202503.0658.v1

normality assumption. Meanwhile, the posteriors of the precision variables of noise σ^{-2}, components α, basis functions β, mean weights η are Gamma distributions:

$$\mathcal{Q}_Z(Z) = \prod_i \mathcal{Q}_{Z_i}(Z_i) = \prod_i \mathcal{N}(Z_i|\mu_{Z_i}, \Sigma_{Z_i}) \tag{6}$$

$$\mathcal{Q}_W(W) = \mathcal{N}(\mathrm{vec}(W)|\mu_{\mathrm{vec}(W)}, \Sigma_{\mathrm{vec}(W)}) \tag{7}$$

$$\mathcal{Q}_{\bar{Z}}(\bar{Z}) = \mathcal{N}(\bar{Z}|\mu_{\bar{Z}}, \Sigma_{\bar{Z}}) \tag{8}$$

$$\mathcal{Q}_\sigma(\sigma) = \Gamma(\sigma^{-2}|a_\sigma, b_\sigma) \tag{9}$$

$$\mathcal{Q}_\alpha(\alpha) = \prod_j \mathcal{Q}_{\alpha_j} = \prod_j \Gamma(\alpha_j|a_{\alpha_j}, b_{\alpha_j}) \tag{10}$$

$$\mathcal{Q}_\beta(\beta) = \prod_k \mathcal{Q}_{\beta_k} = \prod_k \Gamma(\beta_k|a_{\beta_k}, b_{\beta_k}) \tag{11}$$

$$\mathcal{Q}_\eta(\eta) = \Gamma(\eta, |a_\eta, b_\eta) \tag{12}$$

3.1.1. Update Steps

In mean field approximation using the surrogate posterior $\mathcal{Q}_\Theta = \prod_i \mathcal{Q}_{\Theta_i}$ conditioned on observations Y, the lower bound is maximized with respect to each unknown Θ_i. With the conjugate prior, the optimal updates (denoted with "$\leftarrow$") make the moments of $\mathcal{Q}_{\Theta_i}$ equal to the moments conditioned on the remaining parts of $\mathcal{Q}_\Theta$ [72]:

$$\mathcal{Q}_{\Theta_i} \leftarrow \frac{\exp\left(\mathbb{E}_{\mathcal{Q}_{/\Theta_i}}[\ln(\Pr[Y, \Theta])]\right)}{\int \exp\left(\mathbb{E}_{\mathcal{Q}_{/\Theta_i}}[\ln(\Pr[Y, \Theta])]\right) d\Theta_i} \tag{13}$$

From Equation (13), detailed update rules for each variable are presented subsequently, and the derivations of these formulas are in the supplementary material.

Updates for the parameters of the posterior for the precision of components $\mathcal{Q}_{\alpha_j}, \forall j = 1 : J$:

$$a_{\alpha_j} \leftarrow a_0 + \frac{K}{2}, \tag{14}$$

$$b_{\alpha_j} \leftarrow b_0 + \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}_{\mathcal{Q}_{/\alpha_j}}[W_{jk}^2 \beta_k] = b_0 + \frac{1}{2}\sum_{k=1}^{K}\left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right)\frac{a_{\beta_k}}{b_{\beta_k}}\right), \tag{15}$$

where Equation (14) calculates the corrected degrees of freedom and Equation (15) calculates the corrected sum of squares. As a_0 and b_0 approach 0, the expectation of precision α_j, which is $\mathbb{E}_{\mathcal{Q}_{\alpha_j}}[\alpha_j] = \frac{a_{\alpha_j}}{b_{\alpha_j}}$, is exactly the inverse of the empirical or sample variance.

Updates for the parameters of the posterior of the precision of the mean weights $\mathcal{Q}_\eta$:

$$a_\eta \leftarrow a_0 + \frac{K}{2}, \tag{16}$$

$$b_\eta \leftarrow b_0 + \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}_{\mathcal{Q}_{/\eta}}[\bar{Z}_k^2 \beta_k] = b_0 + \frac{1}{2}\sum_{k=1}^{K}\left(\left(\Sigma_{\bar{Z}k} + \mu_{\bar{Z}k}^2\right)\frac{a_{\beta_k}}{b_{\beta_k}}\right) \tag{17}$$

 doi:10.20944/preprints202503.0658.v1

Updates for the parameters of the posterior of the precision of basis functions $\mathcal{Q}_{\beta_k}, \forall k = 1:K$:

$$a_{\beta_k} \leftarrow a_0 + \frac{J+1}{2}, \tag{18}$$

$$\begin{aligned} b_{\beta_k} &\leftarrow b_0 + \frac{1}{2}\mathbb{E}_{\mathcal{Q}_{/\beta_k}}[\bar{Z}_k^2\eta + \sum_{j=1}^{J} W_{jk}^2\alpha_j] \\ &= b_0 + \frac{1}{2}\left(\left(\Sigma_{\bar{Z}kk} + \mu_{\bar{Z}k}^2\right)\frac{a_\eta}{b_\eta} + \sum_{j=1}^{J}\left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right)\frac{a_{\alpha_j}}{b_{\alpha_j}}\right)\right) \end{aligned} \tag{19}$$

Updates for the parameters of the posterior of the mean weights $\mathcal{Q}_{\bar{Z}}$:

$$\Sigma_{\bar{Z}} \leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\sum_{i=1}^{P}\Psi_i + \eta\,\mathrm{diag}(\beta)\right]\right)^{-1} = \left(\frac{a_\sigma}{b_\sigma}\sum_{i=1}^{P}\Psi_i + \frac{a_\eta}{b_\eta}\,\mathrm{diag}(\frac{a}{b})\right)^{-1}, \tag{20}$$

$$\mu_{\bar{Z}} \leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\right]\sum_{i=1}^{P}(Y_i - \mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[Z_i W]\Phi_i)\Phi_i^T\right)\Sigma_{\bar{Z}} = \left(\frac{a_\sigma}{b_\sigma}\sum_{i=1}^{P}(Y_i - \mu_{Z_i}\mu_W\Phi_i)\Phi_i^T\right)\Sigma_{\bar{Z}} \tag{21}$$

where $\mathrm{diag}(\beta)$ denotes the diagonal matrix with diagonal entries given by β. Equation (20) indicates that the eigenvectors of $\Sigma_{\bar{Z}}$ are solely determined by the sum of Gram matrices $\sum_{i=1}^{P}\Psi_i$, while the eigenvalues of $\Sigma_{\bar{Z}}$ have a negative correlation with the scale of $\sum_{i=1}^{P}\Psi_i$, the prior $\eta\,\mathrm{diag}(\beta)$ and data-dependent term σ^{-2}. It is sensible because, for instance, large noise would result in large uncertainty in $\bar{Z}$. In Equation (21), the data residuals, excluding component scores, are projected into the K-dimensional space through the inner product with Φ_i and summed over all sample functions to calculate the mean weights.

Updates for the parameters of the posterior of the weights $\mathcal{Q}_W$:

$$\begin{aligned} \Sigma_{\mathrm{vec}(W)} &\leftarrow \mathbb{E}_{\mathcal{Q}_{/W}}\left[\sigma^{-2}\sum_{i=1}^{P}\left(\Psi_i^T \otimes (Z_i^T Z_i)\right) + \mathrm{diag}(\beta)\otimes\mathrm{diag}(\alpha)\right]^{-1} \\ &= \left(\frac{a_\sigma}{b_\sigma}\sum_{i=1}^{P}\left(\Psi_i^T \otimes ({\mu_{Z_i}}^T\mu_{Z_i} + \Sigma_{Z_i})\right) + \mathrm{diag}\left(\frac{a}{b}\right)\otimes\mathrm{diag}\left(\frac{c}{d}\right)\right)^{-1}, \end{aligned} \tag{22}$$

$$\begin{aligned} \mu_{\mathrm{vec}(W)} &\leftarrow \mathbb{E}_{\mathcal{Q}_{/W}}\left[-\sigma^{-2}\sum_{i=1}^{P}\mathrm{vec}\left(\left(\Phi_i(\Phi_i^T\bar{Z}^T - Y_i^T)Z_i\right)^T\right)^T\right]\Sigma_{\mathrm{vec}(W)} \\ &= -\frac{a_\sigma}{b_\sigma}\sum_{i=1}^{P}\mathrm{vec}\left(\left(\Phi_i(\Phi_i^T{\mu_{\bar{Z}}}^T - Y_i^T)\mu_{Z_i}\right)^T\right)^T\Sigma_{\mathrm{vec}(W)} \end{aligned} \tag{23}$$

Equation (22) is similar to Equation (20), because it is correlated with Φ_i, its prior $\mathrm{diag}(\beta)\otimes\mathrm{diag}(\alpha)$ and data-dependent terms σ^{-2} and Z_i. In Equation (23), the data residual excluding the mean function is used to estimate the expectation of W.

Updates for the parameters of the posterior of the component scores $\mathcal{Q}_{Z_i}$:

$$\begin{aligned} H_{ijk} &\leftarrow \mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_j\Psi_i W_k^T] = \mathrm{Tr}(\mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_k^T W_j]\Psi_i) \\ &= Tr\left(\left(\Sigma_{[W_k,W_j]} + \mu_{[W_j]}^T\mu_{[W_k]}\right)\Psi_i\right), \forall j = 1:K, k = 1:K, \end{aligned} \tag{24}$$

$$\Sigma_{Z_i} \leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/Z_i}}[\sigma^{-2}W\Psi_i W^T + I]\right)^{-1} = [\frac{a_\sigma}{b_\sigma}H_i + I]^{-1}, \tag{25}$$

$$\mu_{Z_i} \leftarrow \mathbb{E}_{\mathcal{Q}_{/Z_i}}[\sigma^{-2}(Y_i - \bar{Z}\Phi_i)\Phi_i^T W^T]\Sigma_{Z_i} = \frac{a_\sigma}{b_\sigma}(Y_i - \mu_{\bar{Z}}\Phi_i)\Phi_i^T(\mu_W)^T\Sigma_{Z_i}, \tag{26}$$

 doi:10.20944/preprints202503.0658.v1

where H_i is a temporary variable denoting the Gram matrix of weighted kernel functions $W\Phi_i$ and $\Sigma_{[W_k,W_j]}$ denotes the covariance between W_k^T and W_j in $\mathcal{Q}$.

Updates for the parameters of the posterior of the noise $\mathcal{Q}_\sigma$:

$$a_\sigma \leftarrow a_0 + \frac{1}{2}\sum_i N_i, \tag{27}$$

$$\begin{aligned} b_\sigma &\leftarrow b_0 + \frac{1}{2}\mathbb{E}_{\mathcal{Q}_{/\sigma}}\left[\sum_i ||Y_i - (Z_i W + \bar{Z})\Phi_i||_2^2\right] \\ &= b_0 + \frac{1}{2}\sum_i (Y_i Y_i^T - 2Y_i(\mu_{Z_i}\mu_W\Phi_i)^T - 2Y_i(\mu_{\bar{Z}}\Phi_i)^T + 2\mu_{Z_i}\mu_W\Psi_i(\mu_{\bar{Z}})^T \\ &+ \mathrm{Tr}\Big(\Big(\Sigma_{\bar{Z}} + (\mu_{\bar{Z}})^T\mu_{\bar{Z}}\Big)\Psi_i\Big)) + \frac{1}{2}\mathrm{vec}(H^T)^T\sum_i \mathrm{vec}\Big(\mathrm{vec}(\Psi_i)\,\mathrm{vec}(\Sigma_{Z_i} + {\mu_{Z_i}}^T\mu_{Z_i})^T\Big), \end{aligned} \tag{28}$$

where H is a temporary variable that is updated by

$$H_{j+kM} \leftarrow \mathbb{E}_{\mathcal{Q}_{/\sigma}}\left[\mathrm{vec}(W_k {W_j}^T)^T\right] = \mathrm{vec}(\Sigma_{[W_k,W_j]} + \mu_{[W_j]}^T\mu_{[W_k]})^T, \forall j = 1:K, k = 1:K \tag{29}$$

Nearly noninformative (vagor) priors, i.e., with almost zero a_0, b_0, introduce an inherent identifiability ambiguity in our formulation, specifically, in the product of the precision parameters α, β and η (Equations 20 and 22). In our model, scaling α and η by a specific factor while inversely scaling β leaves the product (and hence the lower bound in Equation 5) unchanged. This inherent ambiguity can lead α, β, and η to converge to extreme values, thereby challenging numerical stability during optimization. To mitigate this issue, we adopt a heuristic constraint to ensure that the smallest values of α and β remain within one order of magnitude of each other. Specifically, we enforce $\left|\log_{10}^{\left(\frac{\min(\mathrm{ff})}{\min(\beta)}\right)}\right| \leq 1$. If an update to any α_j or β_k would violate this constraint, that particular update is skipped while the rest of the parameters remain updated. This strategy does not alter the algorithm's overall structure but stabilizes the optimization by curbing unnecessary flexibility in the precision parameters.

3.2. *Scalable Update Strategy*

The scalability of our algorithm so far is primarily challenged by the need to optimize the variational lower bound, $\mathcal{L}$ over K basis functions. As indicated by Equation (22), time complexity is $\mathcal{O}(K^6)$ (or, alternatively, $\mathcal{O}(K^2 P \max_i(N_i))$, typically dominated by the former), which becomes prohibitive when K is large. In practice, however, only a small subset of these basis functions is necessary for an accurate representation–those with non-negligible weights under our sparse prior.

To address this, we focus the updates on the subspace of active basis functions, denoted as $K^{(a)}$, which comprises only those functions with non-negligible weights. The remaining basis functions, whose influence is minimal, are held fixed during optimization. Furthermore, the number of active principal components is noted as $J^{(\mathrm{a})}$ and set equal to $K^{(a)}$, ensuring that the model spans the full range of possible ranks from 1 to $K^{(\mathrm{a})}$. Consequently, we optimize $\mathcal{Q}^{(\mathrm{a})}$ using updates derived w.r.t. the objective $K^{(a)}$-dimensional lower bound $\mathcal{L}^{(a)}$ as an efficient surrogate of the full updates of $\mathcal{Q}$ w.r.t. the full lower bound $\mathcal{L}$, using only $K^{(a)}$ active basis functions. Meanwhile, the active dimensionality of the model is adjusted dynamically during optimization by activating or deactivating basis functions based on their precision parameters. For clarity, variables associated with the active subspace are annotated with the superscript (a) (e.g., $a_{\alpha_j}^{(\mathrm{a})} = a_0 + \frac{K^{(\mathrm{a})}}{2}$ versus $a_{\alpha_j} = a_0 + \frac{K}{2}$).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2025 doi:10.20944/preprints202503.0658.v1

3.2.1. Implicit Factorization

For notation clarity, we reorder the rows and columns of our parameter matrices to separate active components from inactive ones. Specifically, we partition as follows:

$$Z_i = \begin{pmatrix} Z_{iA} & Z_{iB} \end{pmatrix}, \bar{Z} = \begin{pmatrix} \bar{Z}_A & \bar{Z}_B \end{pmatrix}, \alpha = \begin{pmatrix} \alpha_A & \alpha_B \end{pmatrix}, \beta = \begin{pmatrix} \beta_A & \beta_B \end{pmatrix},$$
$$W = \begin{pmatrix} W_A & W_B \\ W_C & W_D \end{pmatrix}, \Phi_i = \begin{pmatrix} \Phi_{iA} \\ \Phi_{iB}, \end{pmatrix} \tag{30}$$

Here, the subscript A denotes variables belonging to the active subspace (i.e., those corresponding to $K^{(a)}$ basis functions), while B, C, and D denote the inactive components. Notably, the cross terms W_B and W_C involve both active and inactive components; these are updated implicitly, as proved in the supplements.

Following the strategy in [73], a basis function is deemed inactive if its precision exceeds a high threshold, i.e. $\alpha_j > \epsilon^{-1}$ and $\beta_k > \epsilon^{-1}$ as $\epsilon \to 0$. In the limit, the inactive basis functions decouple from the active ones, leading to the following mean-field factorization:

$$\mathcal{Q}_W = \mathcal{Q}_{W_A} \mathcal{Q}_{W_B} \mathcal{Q}_{W_C} \mathcal{Q}_{W_D} \tag{31}$$
$$\mathcal{Q}_{\bar{Z}} = \mathcal{Q}_{\bar{Z}_A} \mathcal{Q}_{\bar{Z}_B} \tag{32}$$
$$\mathcal{Q}_{Z_i} = \mathcal{Q}_{Z_{iA}} \mathcal{Q}_{Z_{iB}} \, . \tag{33}$$

The factorization of α and β was already obtained in Equations (10) and (11). These factorizations allow us to decouple the update for the active subspace with the proof provided in the supplementary material.

It implies that only updates for $\mathcal{Q}_{Z_{iA}}, \mathcal{Q}_{W_A}, \mathcal{Q}_{W_B}, \mathcal{Q}_{W_C}, \mathcal{Q}_{\bar{Z}_A}, \mathcal{Q}_{\alpha_A}, \mathcal{Q}_{\beta_A}, \mathcal{Q}_\sigma, \mathcal{Q}_\eta$ are required as shown in Figure 2. This strategy reduces the computational complexity from $\mathcal{O}(K^6)$ to $\mathcal{O}({K^{(a)}}^6)$. Moreover, the active dimensions $K^{(a)}$ are initialized using a modified, multi-instance version of relevance vector machine [52], as detailed in Section 11 in the supplementary materials.

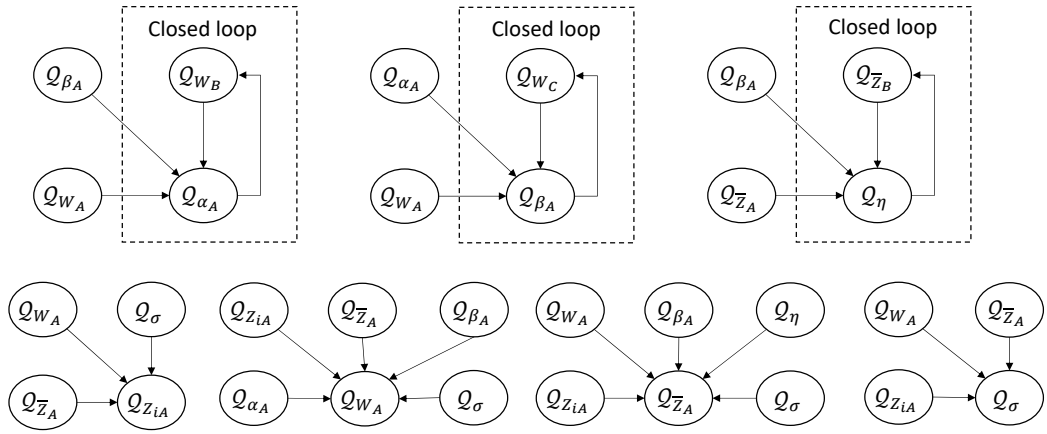


Figure 2. Diagrams of variational inference algorithm for all the parameters. The top 3 diagrams each has a closed loop and a closed form overall transfer function.

3.2.2. Low-dimensional Lower Bound

This section shows how to optimize these active surrogates, e.g., $\mathcal{Q}_{\bar{Z}_A}$, using updates of $\mathcal{Q}^{(a)}$ w.r.t. the $K^{(a)}$-dimensional lower bound $\mathcal{L}^{(a)}$, which ultimately optimizes the full lower bound $\mathcal{L}$. To distinguish between the two, we denote the active surrogate posterior for the full model as $\mathcal{Q}_{\bar{Z}_A}$ and that for the reduced $K^{(a)}$-dimensional model as $\mathcal{Q}^{(a)}_{\bar{Z}_A}$. The active Gaussian surrogate posteriors are

 doi:10.20944/preprints202503.0658.v1

shared, e.g., $\mathcal{Q}_{\bar{Z}_A} = \mathcal{Q}^{(\mathrm{a})}_{\bar{Z}_A} = \mathcal{N}(\bar{Z}_A|\mu_{\bar{Z}A}, \Sigma_{\bar{Z}A})$. This implies updating $\mathcal{Q}^{(\mathrm{a})}$ is equivalent to updating $\mathcal{Q}$, so we set the moments of the active distributions of the full model to match those of the reduced model. However, the surrogate posterior Gamma distributions differ between the two models. For example, the update of $\mathbb{E}_{\mathcal{Q}^{(\mathrm{a})}}[\alpha_A]$ depends solely on $\mathcal{Q}_{W_A}$, whereas $\mathbb{E}_{\mathcal{Q}}[\alpha_A]$ also incorporates a cross term $\mathcal{Q}_{W_B}$ corresponding to the remaining $(K - K^{(a)})$ dimensions. This difference is reflected in how the scale parameters depend on the number of active versus total basis functions, as shown in Equations (14), (16) and (18). Nonetheless, we prove that in the limit $\epsilon \to 0$, the fixed point of the $K^{(a)}$-dimensional updates of the complete surrogate $\mathcal{Q}$ equals that of the reduced surrogate $\mathcal{Q}^{(\mathrm{a})}$. Consequently, the updates for $\mathcal{Q}_{\alpha_A}$, $\mathcal{Q}_{\beta_A}$, and $\mathcal{Q}_\eta$ are derived directly from the expectations of the reduced model $\mathcal{Q}^{(\mathrm{a})}_{\alpha_A}$, $\mathcal{Q}^{(\mathrm{a})}_{\beta_A}$, $\mathcal{Q}^{(\mathrm{a})}_\eta$:

$$\mathbb{E}_{\mathcal{Q}}[\alpha_A] \leftarrow \mathbb{E}_{\mathcal{Q}^{(\mathrm{a})}}[\alpha_A] \Leftrightarrow b_{\alpha_j} \leftarrow \frac{a_{\alpha_j}}{a^{(\mathrm{a})}_{\alpha_j}} b^{(\mathrm{a})}_{\alpha_j}, \forall j \leq J^{(\mathrm{a})}, \tag{34}$$

$$\mathbb{E}_{\mathcal{Q}}[\beta_A] \leftarrow \mathbb{E}_{\mathcal{Q}^{(\mathrm{a})}}[\beta_A] \Leftrightarrow b_{\beta_k} \leftarrow \frac{a_{\beta_k}}{b_{\beta_k}} b^{(\mathrm{a})}_{\beta_k}, \forall k \leq K^{(\mathrm{a})}, \tag{35}$$

$$\mathbb{E}_{\mathcal{Q}}[\eta] \leftarrow \mathbb{E}_{\mathcal{Q}^{(\mathrm{a})}}[\eta] \Leftrightarrow b_\eta \leftarrow \frac{b^{(\mathrm{a})}_\eta}{a^{(\mathrm{a})}_\eta} a_\eta \tag{36}$$

These update Equations (34), (35), (36) prove to optimize $\mathcal{L}$ in Theorem 1, 2 in the supplements.

3.2.3. Heuristic for Activating Basis Functions

The proposed method selects a relatively small set of basis functions from a potentially extensive set of possibilities. The computational costs are kept in check by recognizing that inactive basis functions do not interact with those active (with non-negligible weights). Due to computational constraints, we consider functions for activation sequentially rather than all at once. Thus, we propose Algorithm 1 to introduce unseen basis functions into the active set using a selective strategy akin to the heuristic approach described in [73].

The algorithm selects the top function, ϕ_{Bk}, from the inactive basis functions $\{\phi_{Bk}\}_k$ by gauging their correlation with residuals and applying an angle-based threshold τ_{ang} relative to the subspace of ϕ_A. The correlation with residuals for ϕ_{Bk} is measured by $\sum_i \left(\Phi_{iBk}(Y_i - \mathbb{E}_{\mathcal{Q}^{(\mathrm{a})}}[Z_{iA}W_A + \bar{Z}_A]\Phi_{iA})^T\right)^2$. The angle-based threshold ensures a meaningful distinction from active functions. Next, the current active surrogate posterior is expanded by a dimension for $\tilde{\phi_{Bk}}$, initiating optimization from the numerical maximum $\tau_{\max}$. Post optimization, the function gets retained if it falls below $\tau_{\max}$. Otherwise, the algorithm terminates. Efficiently, in trial optimization, the approach replaces one function with precision $\tau_{\max}$, if present.

Algorithm 1 Search for new basis functions to activate

Sort inactive basis functions $\{\phi_{Bk}\}_k$ by correlation with residuals.
Filter through $\{\phi_{Bk}\}_k$, selecting the most correlated one as ϕ_{Bk}.
Copy current active surrogate $\mathcal{Q}^{(\mathrm{a})}(\Theta)$ posterior to $\mathcal{Q}_k{}^{(a)}(\Theta)$.
Expand dimension in $\mathcal{Q}_k{}^{(a)}(\Theta)$ for ϕ_{Bk}.
Optimize $\mathcal{Q}_k{}^{(a)}(\Theta)$ for 3 iterations using mean field approximation.
if expected precision is within threshold **then**
    $\mathcal{Q}^{(\mathrm{a})}(\Theta) \leftarrow \mathcal{Q}_k{}^{(a)}(\Theta)$.
end if

4. Faster Variant

To enhance the computational efficiency of our primary algorithm, we introduce a faster variant, denoted as BSFDA$^{\text{Fast}}$. This approach leverages conditional independence among the columns of W, enabling separate updates and thereby reducing computational complexity. Similar strategies have

 doi:10.20944/preprints202503.0658.v1

been described in [34,62]. The model is defined with an introduced variable ζ_i for the coefficient noise as follows:

$$\theta_i = Z_i W + \zeta_i, \tag{37}$$

$$\zeta_{ik} \sim \mathcal{N}(0, \varsigma_k^2 \beta_k^{-1}). \tag{38}$$

Similar to before, we assign a conjugate Gamma prior to the precision:

$$\varsigma_k^{-2} \sim \Gamma(a_0, b_0). \tag{39}$$

This formulation ensures that the columns of W are conditionally independent, allowing the variational distribution to factorize as: $\mathcal{Q}_{\mathcal{W}} = \prod_k \mathcal{Q}_{W_{\cdot k}}$, thereby facilitating separate updates for each column. Consequently, the time complexity is reduced from $\mathcal{O}({K^{(a)}}^6)$ to $\mathcal{O}({K^{(a)}}^3)$.

To align with the original model, it is necessary for ζ and the associated variance parameters ς to approach zero. Having ς too high would allow the coefficient noise to corrupt the signal, biasing the model toward underestimating the true signal levels, particularly because this noise operates in the coefficient space where it introduces smooth, correlated variation (low entropy, like signals) that is harder to eliminate than high-frequency white noise (maximum entropy). Injecting the same amount of noise leads to an unbiased estimation of the signals but increases the estimation variance. Conversely, as ς decreases, the columns of W become dependent, violating the independence assumption inherent in variational inference. This dependency degrades the approximation quality and slows down the optimization process. Such dependency issues are well documented in both variational inference and MCMC literature–with recent efforts addressing them via structured VI [72] or blocked/collapsed Gibbs sampling [74]. Empirical validations of this noise impact are conducted with both BSFDA$^{\text{Fast}}$ in Section 5 and with Bayesian PCA [45,62] in Section 12.2 in the supplements.

To balance the trade-off between optimization speed and accuracy, we adopt a strategy of gradually decreasing the values of ς_k during the optimization iterations. Specifically, we initialize ς_k with a relatively large value and linearly decrease it from 10^{-2} to 10^{-5} over the first half of the iterations. After reaching 10^{-5}, ς_k is fixed for the remaining iterations. This gradual reduction ensures that the algorithm initially maintains efficiency with benefits from minimizing interdependency among the columns of W to accelerate convergence while later preserving quality of the approximation by preventing the noise from obscuring signal components. We unify the scales by scaling the basis functions so that Z_i is standard normal and W is an identity matrix in initialization. Empirical evaluations indicate the strategy above is effective in most applications.

By implementing these modifications, BSFDA$^{\text{Fast}}$ offers a practical solution that substantially accelerates the algorithm without significant loss in accuracy, making it well-suited for large-scale, high-dimensional functional data analysis.

5. Results

The proposed method proves its effectiveness through simulations and applications to observed data sets.

5.1. Simulation Results

In simulations, we evaluate functional data analysis performance in model selection, estimated covariance accuracy, and extendability to multi-dimensional domains.

The model selection metric is the accuracy in estimating the number of principal components, which is the dimension of the compact subspace of signal variations. The configuration of the simulations in this section aligns with that established in [37], covering various scenarios. Simulated data sets derive from a latent generative model with variables $\mathcal{Z}_i$ with dimension r for the i-th sample function and noise corruption with a standard deviation of σ: $Y_i = \sum_{j=1}^{r} \big(\mathcal{Z}_{ij} f_j(X_i)\big) + g(X_i) +$

 doi:10.20944/preprints202503.0658.v1

$E_i, \mathcal{Z}_{ij} \sim \mathcal{N}(0, v_j), E_j \sim \mathcal{N}(0, \sigma^2 I)$, where $\{f_j\}_{j=1}^r$ represent eigenfunctions, $\{v_j\}_{j=1}^r$ are the eigenvalues, $g : R \mapsto R$ signifies the mean function. Here, we consider five scenarios.

Scenario 1: Data generated with $g = 5(x - 0.6)^2$, $r = 3$, $v = (0.6, 0.3, 0.1)$, $\sigma^2 = 0.2$, $f_1(x) = 1, f_2(x) = \sqrt{2}\sin(2\pi x), f_3(x) = \sqrt{2}\cos(2\pi x)$. Here $v_3 < \sigma^2$, i.e., the noise has a larger variance than the smallest signal.

Scenario 2: Similar to Scenario 1, but the third eigenfunction is replaced by a function with higher frequencies $f_3(x) = \sqrt{2}\cos(4\pi x)$, and the principal component scores follow a skewed Gaussian mixture model. Specifically, the j-th component score has $1/3$ probability of following a $\mathcal{N}(2\sqrt{v_j/3}, v_j/3)$ distribution, and $2/3$ probability of following $\mathcal{N}(-\sqrt{v_j/3}, v_j)$, for $j = 1, 2, 3$.

Scenario 3: Data generated with $g = 12.5(x - 0.5)^2 - 1.25$, $r = 3$, $v = (4, 2, 1)$, $\sigma^2 = 0.5$, $f_1(x) = 1, f_2(x) = \sqrt{2}\cos(2\pi x), f_3(x) = \sqrt{2}\sin(4\pi x)$.

Scenario 4: Same as Scenario 3, but the component scores are generated from a Gaussian mixture model as Scenario 2.

Scenario 5: Data from $g = 12.5(x - 0.5)^2 - 1.25$, $r = 6$, $v = (4, 3.5, 3, 2.5, 2, 1.5)$, $\sigma^2 = 0.5$, $f_1(x) = 1, f_{2k}(x) = \sqrt{2}\sin(2k\pi x)$ for $k = 1, 2, 3, f_{2k+1}(x) = \sqrt{2}\cos(2k\pi x)$ for $k = 1, 2$, j-th component score obeying $\mathcal{N}(0, v_j)$.

In each scenario, simulations produced 200 sample functions. We investigated 3 cases with sparse, medium, and dense sampling by assigning the number of observations per sample function $N_i = \{5, 10, 50\}$. Each case in each scenario is repeated 200 times. The method's performance was compared to *fpca* from [22], AIC and BIC in the 2022 release of *pace* [20], modified AIC and BIC in [37], and all the competing methods in [37]. For *fpca*, we set the candidate numbers of basis functions as [8,10,15,20], and the candidate dimensions of the process as [2,3,4,5] for Scenario 1-4 and [4,5,6,7,8] for Scenario 5. The other parameters are all set to the defaults. Due to its consistent overestimation of the true number of components–likely resulting from interference by correlated noise and less sparse precision priors–we excluded LFRM[34] from further comparisons (see Section 12.1.1 in the supplements).

Each estimation chose ten length-scales of functions, which are selected using cross-validation and k-means clustering. This adaptive strategy allows the algorithm to choose distinct length-scales at different locations of the definition domain, thereby accommodating varying smoothness characteristics inherent in complex functional data–a flexibility that is not possible when using a regular grid that forces a single length-scale across the entire domain [34]. Sparse sampling in Scenario 5 used five length-scales to avoid over-fitting. Figure 3 shows the length-scales and centers of the selected kernel basis functions for three different numbers of sample points, N_i, in a random repetition of Scenario 5. The results reveal that the selected length-scales mainly concentrate around 0.07, with a few as high as 0.35–suggesting that the lower length-scales capture finer, high-frequency variations. The higher length-scales model the overall, lower-frequency quadratic mean structure and the constant baseline component. Furthermore, the estimated density functions of the selected length-scales exhibit consistent patterns across the three sampling densities, and the method selects 9, 11, and 12 basis functions respectively, demonstrating the algorithm's adaptive fidelity and complexity based on the available observations. The supplements showcase the uncertainty evaluation in Figure 12.

 doi:10.20944/preprints202503.0658.v1

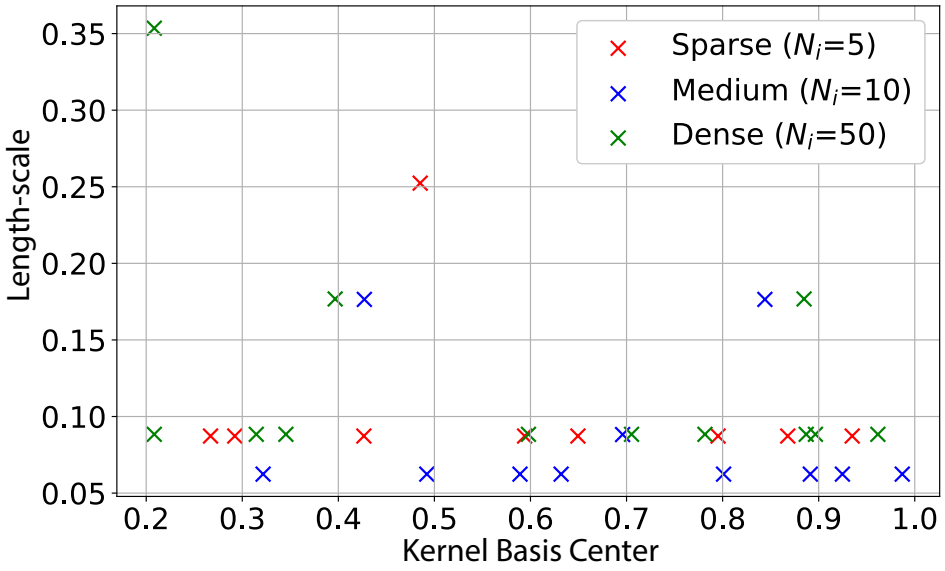


Figure 3. length-scales and centers of selected kernel basis functions in a random repetition for three different m values in Scenario 5.

Tables 1, 2, 3, 4 and 5 show the results. Results for the first five methods are from [37]. Out of 15 cases, the proposed *BSFDA* exhibits the highest accuracy in 12. In the other 3 cases, the accuracy of *BSFDA* is comparable to the best result and is always above 0.950. BSFDA$^{\text{Fast}}$ demonstrates performance comparable to *BSFDA* when applied to medium-density and dense datasets with significantly higher efficiency which we detail in Figure 5 later. However, its efficacy diminishes with sparse data. This limitation arises because the parameter ς can bias model estimation in scenarios with insufficient data evidence, leading to an underestimation of signal variance. Consequently, BSFDA$^{\text{Fast}}$ tends to underestimate the number of components, particularly those capturing nuanced variations, in the presence of sparse observations. Nonetheless, with adequate data, BSFDA$^{\text{Fast}}$ achieves performance on par with the original model.

Table 1. Proportion of accurate estimations for Scenario 1 (r=3).

N_i	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	$\text{AIC}^{2022}_{\text{PACE}}$	$\text{BIC}^{2022}_{\text{PACE}}$	fpca	BSFDA	BSFDA$^{\text{Fast}}$
5	0.000	0.580	0.380	0.410	0.735	0.650	0.880	0.645	**0.995**	0.015
10	0.000	0.980	0.670	0.955	0.985	0.880	0.920	0.645	**1.000**	0.910
50	0.000	**1.000**	0.830	**1.000**	**1.000**	**1.000**	**1.000**	0.890	0.980	0.945

Table 2. Proportion of accurate estimations for Scenario 2 (r=3).

N_i	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	$\text{AIC}^{2022}_{\text{PACE}}$	$\text{BIC}^{2022}_{\text{PACE}}$	fpca	BSFDA	BSFDA$^{\text{Fast}}$
5	0.005	0.630	0.245	0.375	0.605	0.570	0.620	0.475	**1.000**	0.040
10	0.000	0.710	0.665	0.570	0.805	0.825	0.850	0.640	**1.000**	0.995
50	0.000	0.630	0.795	0.955	0.945	**1.000**	**1.000**	0.950	**1.000**	0.950

 doi:10.20944/preprints202503.0658.v1

Table 3. Proportion of accurate estimations for Scenario 3 (r=3).

N_i	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	$\text{AIC}^{2022}_{\text{PACE}}$	$\text{BIC}^{2022}_{\text{PACE}}$	fpca	BSFDA	$\text{BSFDA}^{\text{Fast}}$
5	0.005	0.720	0.325	0.640	0.590	0.320	0.400	0.450	**0.995**	0.945
10	0.000	0.580	0.770	0.965	0.665	0.740	0.755	0.440	0.995	**1.000**
50	0.000	1.000	0.775	1.000	1.000	**1.000**	**1.000**	0.765	0.980	0.920

Table 4. Proportion of accurate estimations for Scenario 4 (r=3).

N_i	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	$\text{AIC}^{2022}_{\text{PACE}}$	$\text{BIC}^{2022}_{\text{PACE}}$	fpca	BSFDA	$\text{BSFDA}^{\text{Fast}}$
5	0.015	0.710	0.410	0.640	0.560	0.515	0.575	0.370	**1.000**	0.975
10	0.000	0.830	0.775	0.920	0.900	0.750	0.760	0.350	**0.995**	0.990
50	0.000	0.945	0.835	**1.000**	**1.000**	**1.000**	**1.000**	0.730	0.950	0.935

Table 5. Proportion of accurate estimations for Scenario 5 (r=6).

N_i	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	$\text{AIC}^{2022}_{\text{PACE}}$	$\text{BIC}^{2022}_{\text{PACE}}$	fpca	BSFDA	$\text{BSFDA}^{\text{Fast}}$
5	0.705	0.470	0.090	0.070	0.545	0.425	0.410	0.855	**0.925**	0.160
10	0.065	0.570	0.525	0.775	0.705	0.575	0.575	0.500	**1.000**	0.930
50	0.000	0.260	0.590	0.980	0.965	0.870	0.770	0.695	**0.995**	0.925

5.1.1. Mean Squared Error in Covariance Operator

The mean squared error across X_{grid}, a grid of 1000 index points:

$$\frac{||\operatorname{cov}(X_{\text{grid}}, X_{\text{grid}}) - \hat{\operatorname{cov}}(X_{\text{grid}}, X_{\text{grid}})||_F^2}{1000 \times 1000},$$

where $||\cdot||_F$ is the Frobenius norm, measure the accuracy of the estimated covariance. The quadratic measure of error with Frobenius norm for covariance estimators has been used by [75]. Methods compared include *fpca* of [22], *pace* of [20] with AIC and BIC, *refund-sc* of [21]. Only cases in scenario 5 were used because of the time constraints (e.g., refund-sc takes 6 hours for 20 repetitions with 50 points in scenario 5). As the most challenging, scenario 5 should provide the most compelling comparison. The results in Table 6 demonstrate that the proposed method is comparable to the best work in terms of estimated covariance accuracy. Specifically, dense sampling becomes prohibitive for *refund-sc*. The results highlight the benefit of continuous formulations, as seen in both *fpca* and the proposed method, over the grid-based optimization in conventional methods. $\text{BSFDA}^{\text{Fast}}$ again performed comparably well given data was adequate.

Table 6. Mean squared error of covariance $\text{Error}_{\text{CovFunc}}$ for Scenario 5

N_i	$\text{AIC}^{2022}_{\text{PACE}}$	$\text{BIC}^{2022}_{\text{PACE}}$	fpca	refund.sc	BSFDA	$\text{BSFDA}^{\text{Fast}}$
5	12.373 ± 4.026	12.377 ± 4.031	**5.192 ± 6.166**	8.833 ± 4.730	5.814 ± 3.535	10.292±12.717
10	10.391 ± 2.521	10.391 ± 2.521	2.098 ± 1.425	5.314 ± 3.501	**2.068 ± 1.427**	2.656±1.712
50	9.054 ± 1.683	9.054 ± 1.683	1.642 ± 1.240	N/A	**1.638 ± 1.247**	1.770±1.275

 doi:10.20944/preprints202503.0658.v1

5.1.2. Multidimensional Functional Data Simulation

A simulation experiment with a four-dimensional index set reveals the proposed method's advantages for high-dimensional data, where the gridding strategies of previous methods are impractical. The settings are as follows with a length-scale $l_s = 0.33$:

$$Z_i \sim \mathcal{N}(0, I) \in R^{1\times 3} \tag{40}$$

$$\phi_0(x) = (\pi l_s^2)^{-2} \exp\left(-\frac{1}{2}\left\|\frac{x - [0.5, 0.5, 0.5, 0.5]}{l_s}\right\|_2^2\right) \tag{41}$$

$$\phi_1(x) = (\pi l_s^2)^{-2} \exp\left(-\frac{1}{2}\left\|\frac{x - [0.4, 0.4, 0.4, 0.4]}{l_s}\right\|_2^2\right) \tag{42}$$

$$\phi_2(x) = (\pi l_s^2)^{-2} \exp\left(-\frac{1}{2}\left\|\frac{x - [0.6, 0.6, 0.6, 0.6]}{l_s}\right\|_2^2\right) \tag{43}$$

$$y_i(x) = Z_{i0} * \sqrt{0.6} * (\phi_0(x) - \phi_1(x)) + Z_{i1} * \sqrt{0.3} * \phi_1(x) + Z_{i2} * \sqrt{0.4} * \phi_2(x) \tag{44}$$

The observations include additive noise with a sigma of 4.472e-01. The cross-validation selects a length-scale of 0.405. The estimated noise sigma is 4.637e-01. The proposed method correctly estimates the number of principal components as 3 and selected 31 basis functions. As shown in Figure 4, the eigenfunctions are correctly estimated. In addition, the estimated mean function is zero, which is accurate.

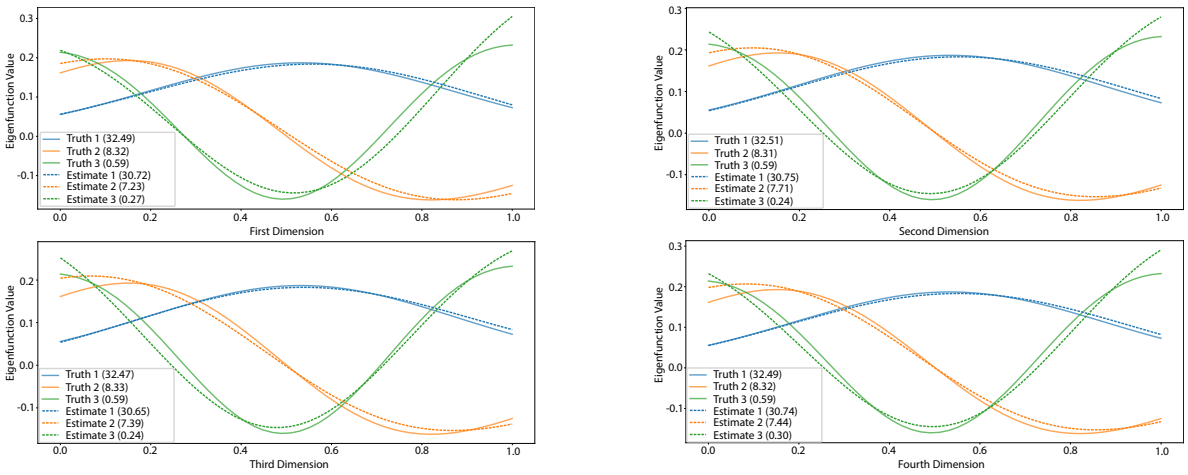


Figure 4. Cross-sectional visualization of eigenfunctions (eigenvalues) of the 4D simulation.

Next, we present a convergence comparison between BSFDA and BSFDA$^{\text{Fast}}$ under four different schedules for the coefficient noise ς_k. Specifically, we compare the default diminishing schedule from 10^{-2} to 10^{-5} with three fixed settings: 10^{-2}, 10^{-3}, and 10^{-5}. We evaluate the covariance error and the discrepancy between the estimated/true dimensionality in one replicate of each sample density in Scenario 5, and the 4D simulation. For the 4D, we adopt a default initial ς_k of 10^{-3}. As illustrated in Figure 5, BSFDA$^{\text{Fast}}$ achieves comparable accuracy to BSFDA while converging significantly faster than BSFDA in terms of both covariance errors and component estimation for medium and densely sampled data. In the 4D case, BSFDA$^{\text{Fast}}$ converges in covariance estimation after approximately 10,000 seconds and in dimensionality after around 4,000 seconds, compared to roughly 100,000 seconds and 13,000 seconds, respectively, for BSFDA. However, for sparse data, BSFDA$^{\text{Fast}}$ exhibits reduced estimation accuracy and underestimates the number of components by one. A similar decline in accuracy is observed in the 4D simulation when data sparsity is high. This limitation arises because the introduction of coefficient noise ς biases the model towards eliminating signals that are deemed insignificant. Moreover, when comparing the three fixed-ς_k variants of the fast algorithm, a clear

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2025 doi:10.20944/preprints202503.0658.v1

trade-off emerges: smaller ς_k reduce overall error but slow down the optimization due to increased dependency among variables. These results collectively demonstrate the effectiveness of our chosen ς_k schedule in BSFDA$^{\text{Fast}}$, as it balances both efficiency and accuracy.

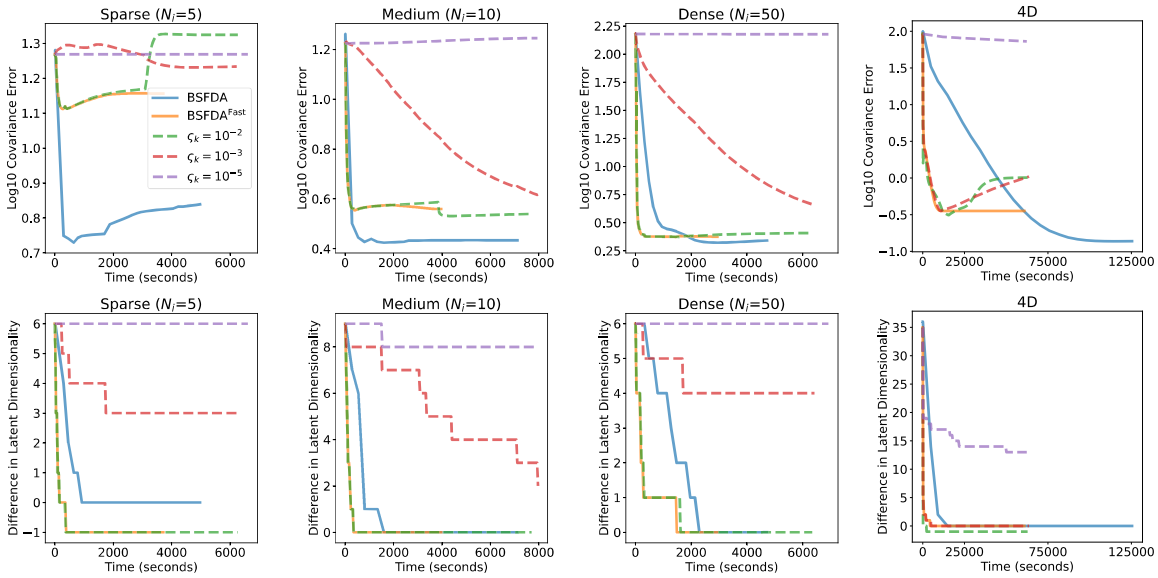


Figure 5. Convergence plots for Scenario 5 in Yehua and the 4D simulation. The upper row displays the covariance error against time, while the lower row illustrates the difference between the estimated and true number of components.

5.2. Results on Public Data Sets

The proposed method's practicality was validated with 2 application data sets, CD4 and wind speed measurements.

5.2.1. CD4

CD4 data, a classical form of functional data, received attention in [1,20,22]. CD4 cell counts gauge the immune system's response to human immunodeficiency virus (HIV) infection, which leads to a progressive reduction in CD4 cell counts. The Multicenter AIDS Cohort Study (MACS) [76] provided the CD4 data. This dataset consists of CD4 percentages from 283 male human subjects that were HIV positive, each with 1 to 14 repeated measurements over time in years. Subjects were scheduled for reevaluation at least semiannually. However, missed visits caused a sparse and uneven distribution of measurements. The proposed method used five length-scales selected from cross-validation and k-means clustering. Finally, the model selected 9 basis functions. Figure 6 displays the estimated mean function, eigenfunctions, and curves of the observations. The mean function reflects the overall decreasing tendency with the progression of the disease. The eigenfunctions are obtained by applying singular value decomposition of the covariance operator that is discretized (for visualization purposes only) with a grid of 50 evenly spaced points over the whole timeline. The first eigenfunction is relatively flat and mainly captures the subject-specific average magnitude of the CD4 counts, consistent with the finding of [1,20,22]. The second eigenfunction captures the simple linear trend of the variations, as described in [22]. The third eigenfunction captures the piece-wise linear time trend with a breakpoint near 2.5 years since baseline. [1,20] found similar eigenfunctions.

 doi:10.20944/preprints202503.0658.v1

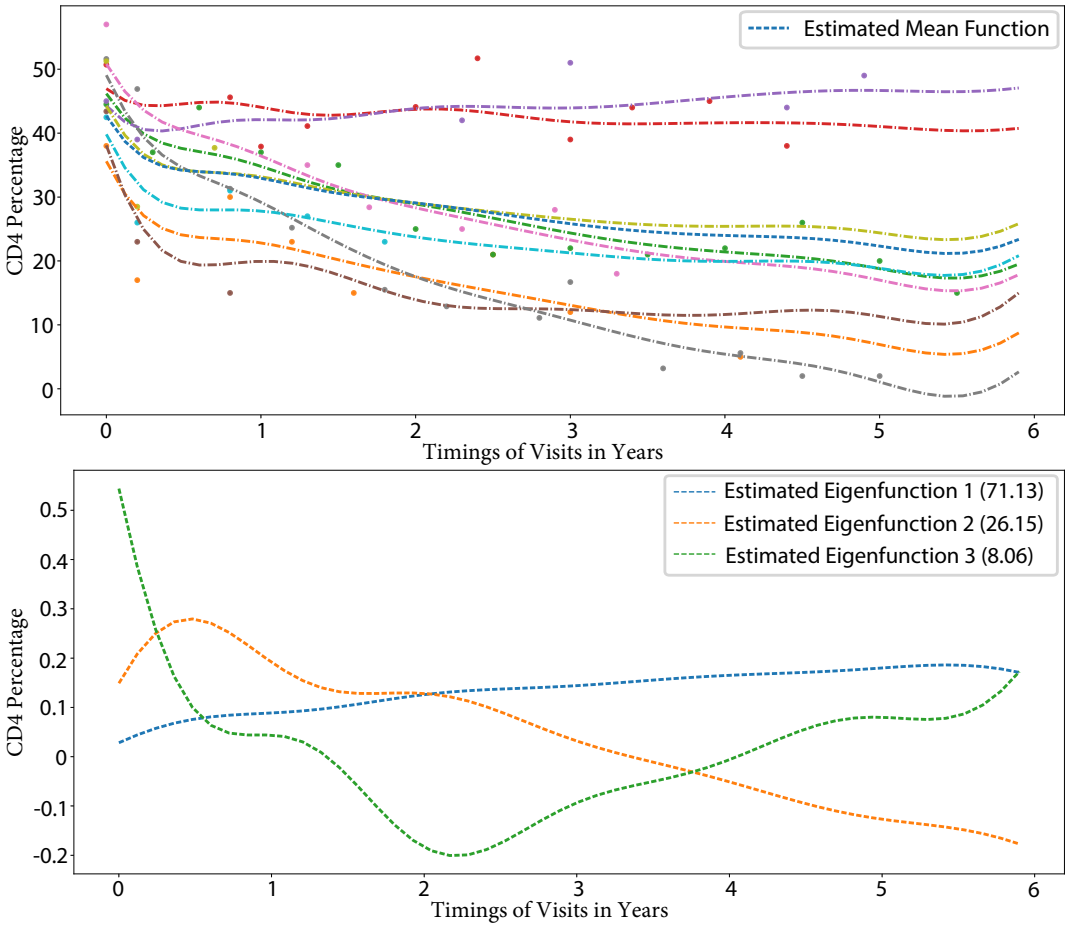


Figure 6. Outcomes from the proposed method applied to MACS CD4 data sets.(top): Estimated curves for a random selection of 9 sampled functions and mean function; (bottom): Estimated eigenfunctions (eigenvalues).

5.2.2. Wind Speed

Wind-speed data, collected from 110 locations across Utah's Salt Lake Valley, varies between 11 to 1440 measurements. The proposed method leverages ten length-scales selected from cross-validation and k-means clustering. Figure 7 illustrates the estimated mean function, curves of the observations, eigenfunctions, and covariance. The horizontal axis represents the seconds starting from 12:00 AM Greenwich Mean Time (GMT) on June 15, 2023, which corresponds to 6:00 PM in Salt Lake City. In Figure 7a, the estimated mean function depicts two pronounced peaks observed approximately at 8:00 PM and 6:00 AM, as well as two troughs around 12:00 AM and 12:00 PM. This pattern aligns with the diurnal cycle, particularly highlighting the thermal activities associated with sunset and sunrise. The peaks during sunset and sunrise are due to the interplay of topographical features, which result in specific breezes, such as the land breeze near the Great Salt Lake and the distinct mountain and valley breezes. The troughs, on the other hand, reflect moments when the atmosphere is at its most stable, with minimal thermal activities disrupting wind patterns. The complexity of the data is distilled and represented using 12 descriptors with 17 basis functions. As Figure 7b shows, the primary eigenfunction is relatively level, indicating that the most significant variation is the location-specific average magnitude. Its profile echoes the influence of sunrise and sunset observed in the mean function, with elevations around 7:00 PM and 5:00 AM and subdued patterns during other times, indicative of a similar atmospheric stability. The estimated covariance in Figure 7c highlights variance peaks around 8 PM and 5 AM, as well as a strong correlation between these periods. This underscores the effects of location-specific topographic factors on wind speed.

 doi:10.20944/preprints202503.0658.v1

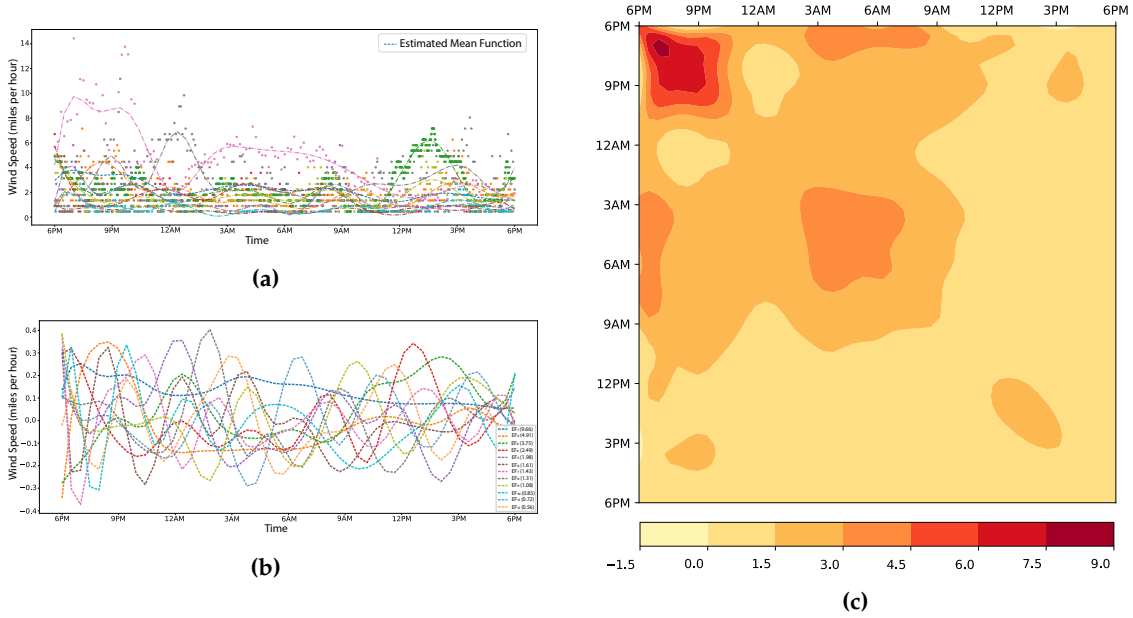


Figure 7. Outcomes from the proposed method applied to a wind speed data set. (a) Estimated curves for a random selection of 9 sampled functions and mean function; (b) Estimated eigenfunctions (eigenvalues) denoted as EF; (c) Estimated covariance.

5.2.3. Modeling Large-Scale, Dynamic, Geospatial Data

Here, we demonstrate the scalability on both the size of the measurements and the dimensionality of our framework. For this, we apply it to the ARGO dataset, which consists of ocean temperature measurements from more than 4,000 locations, at multiple depths, and time points [71]. ARGO is a nearly global observing system for ocean temperature, salinity, and other key variables via autonomous profiling floats. As of 2019, ARGO has generated over 338 gigabytes of data from 15,231 floats [71]. We focused on *high-quality* ("research" mode option in the database API) data from 1998 to 2024 for depths between 0–200 meters in the open-access snapshot of Argo GDAC of November 9, 2024 [77]. The number of measurement points per year varies widely–from 38,931 up to over 11 million, with 127 million in total. Figure 8 illustrates a global map of sea surface temperature measurements from February 2021, highlighting the dataset's extensive spatial coverage.

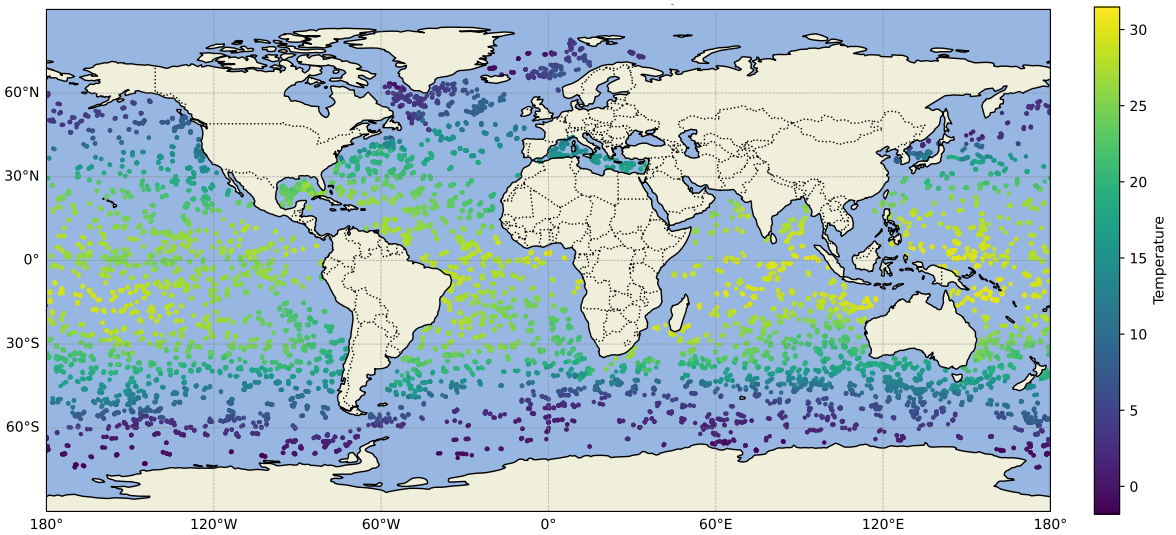


Figure 8. Temperature measurements in 2021 February near the sea surface in the ARGO dataset.

 doi:10.20944/preprints202503.0658.v1

In our modeling, each year's data is treated as a single underlying function of four variables: latitude, longitude (on the spherical Earth), depth, and intra-annual time (modeled as a periodic variable). Note that the spatial data lies on a sphere and the time is a circle, assuming the periodicity of the time of the year. Our approach models these measurements holistically–without resorting to moving windows or sub-modeling–thereby preserving the continuous nature of the data and enabling the extraction of meaningful global, seasonal, and depth-dependent trends. Furthermore, the unique geospatial and temporal structure of Argo data, with spatial coordinates on a sphere and time exhibiting periodicity, necessitates specialized modeling techniques. Given that our model is 4 dimensional, the 4D kernel is defined as a product of the following kernels, following the design strategy for climatological data in [19]. The geospatial kernel on the sphere is a radial basis function (RBF) on geodesic distances. To ensure periodicity, the temporal kernel is an Exp-Sine-Squared $k(x, x') = \exp\left(-\frac{2\sin^2(\pi|x-x'|)}{l_s}\right)$ where l_s is the length-scale. For depth, we use a Gaussian kernel.

The numeric data (excluding metadata) as input to the model was approximately 4 GB. For length-scale selection, we used Gaussian process regression on a small subset of 2,000 randomly selected data in 2016 (medium size of measurements) for a cross validated RMSE which we optimize with a grid search. The specific length scales were set as follows: geodesic length scale of 2×10^3 km, depth length scale of 70 m, time length scale of 3, and periodicity of 1. For evaluation, we held out 10% of the depth profiles (a single round trip of a buoy from surface to a depth at the same coordinate) from each year as testing data, following [78]. The total training set contained roughly 114 million points. Because the sample spacing is typically small relative to the selected length-scales, we apply agglomerative clustering to 10,000 randomly chosen index points, reducing them to 2,000 candidate basis functions. These candidate basis functions–precomputed for efficiency–took roughly 1.7 TB of memory. Computations were performed with 24 threads on a server equipped with 192 Intel® Xeon® Platinum 8360H CPUs @ 3.00GHz and 3TB RAM. Initialization was conducted using the modified RVM for 200 iterations for initial basis functions, using a stochastic optimization with a 1,000-batch size per year. Then BSFDA$^{\text{Fast}}$ executed for 10,000 iterations, where the heuristic to include new bases also used a 1,000-batch size per year. With these computational strategies and heuristics, the entire modeling process was completed in 15 hours.

The proposed approach selected 163 effective basis functions and condensed them into 16 principle components. The final model occupies merely 50 MB of storage. The interpolation yielded a root mean square error (RMSE) of 1.95 and an R^2 of 94.2% on testing data, reflecting a reasonable balance between global dimension reduction and fidelity. The estimated white noise level was also 1.95, indicating that the training data adequately covers the underlying variability in the ARGO observations, and the final model is reasonably generalizable.

Figure 9 presents 2D visualizations of geospatial interpolations at three depths (in decibars, roughly meters) and a specific time (May 29, 2021) around 1°S and 30°W, each with three views. We have picked one measurement as the central point, denoted as the red circle, and selected a narrow window (±1 decibar, ±1 day) around this center. The cyan and fuchsia circles represent training and testing data, respectively, within this window. Their sizes indicate distance along the unplotted dimensions (depth and time here), reflecting variations in these dimensions. The visualizations show that temperatures are warmer near the equator and decrease with depth. The match between interpolated values and actual measurements demonstrates consistency in capturing broad spatial and vertical variations.

 doi:10.20944/preprints202503.0658.v1

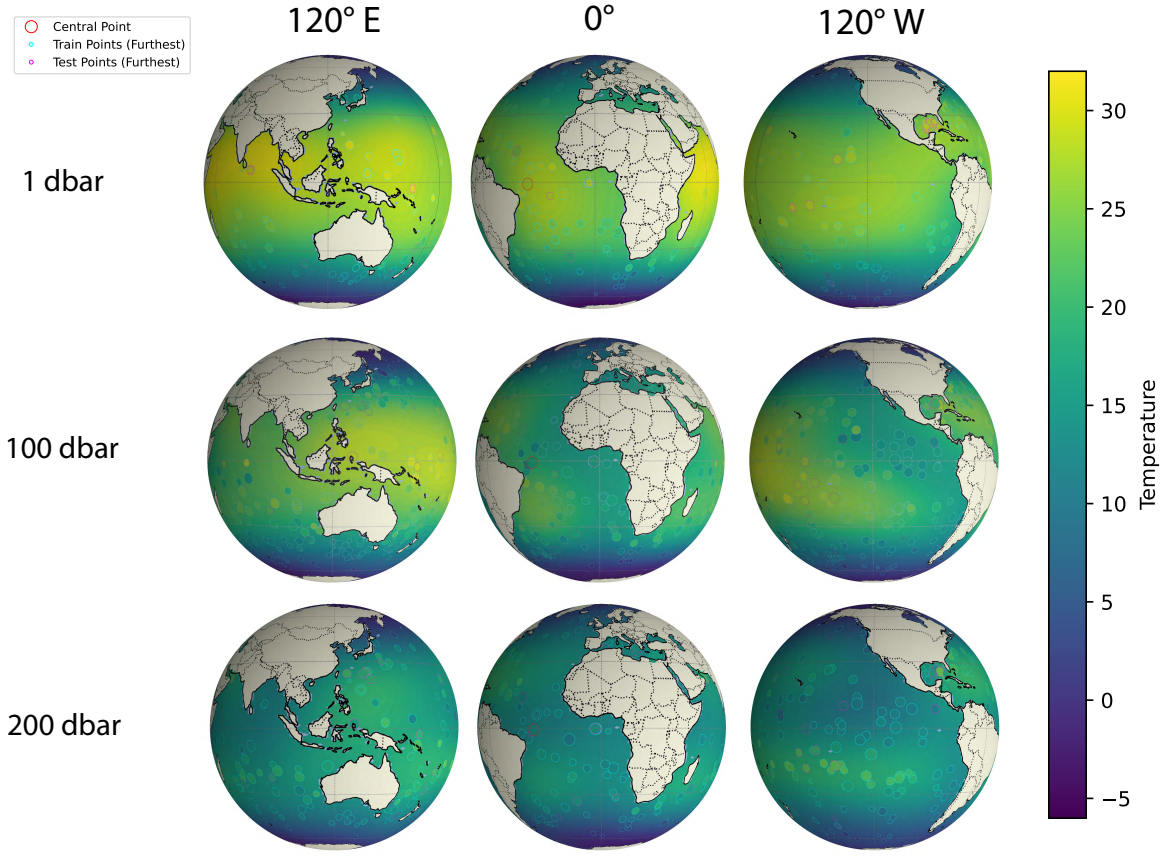
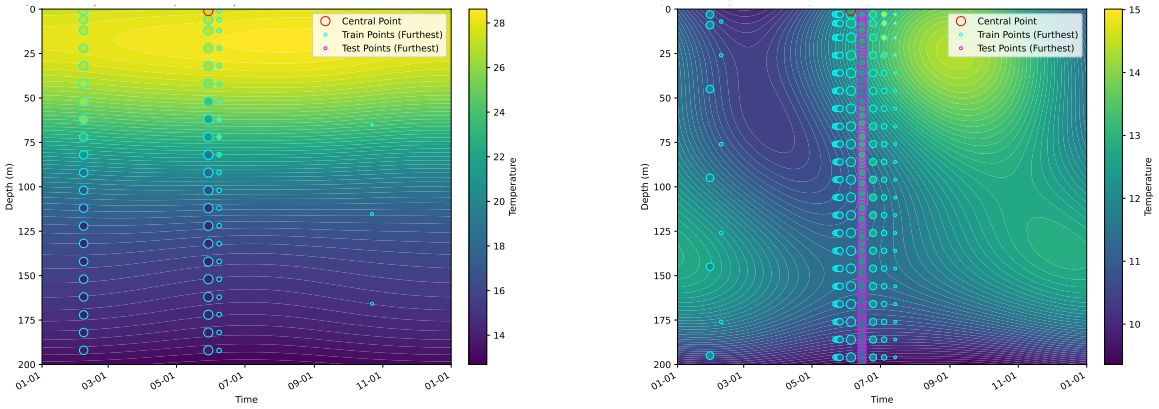


Figure 9. Geodesic interpolation from BSFDA$^{\text{Fast}}$ vs. actual ARGO global oceanic measurements at 1, 200, and 300 decibars, at 1°S and 30°W on May 29. Measurements are represented by circles, with filling color indicating temperature. Circle sizes show distance in depth and time from the central point.

Figure 10 complements this by illustrating interpolation in the depth-time slices while holding the geospatial coordinates fixed, focusing on mixed layer characteristics. The "mixed layer" refers to a region of nearly uniform temperature, which is crucial for understanding thermodynamic potential and nutrient cycling [79]. Here, the plot uses a window of 50 km to include actual measurement, and the circle sizes denote geodesic distance from the chosen center. We plot every fifth measurement vertically to reduce overlap and improve clarity. Figure 10a uses the same center point, 1°S and 30°W, as in Figure 9, exhibiting a shallow mixed layer with pronounced vertical gradients. In contrast, Figure 10b adopts a center at a higher latitude, 49°N and 29°W, where the model reveals a deeper mixed layer. The temperature there remains relatively stable below the surface. The dominant variations are cyclic seasonal changes, which are warmer near the surface around September. As is shown, the vertical sequence of the center and the nearby testing sequence match the interpolation closely. These results confirm that the mid-latitudes exhibit a stronger seasonal cycle [79], and that BSFDA$^{\text{Fast}}$ accurately approximates the actual measurements.

 doi:10.20944/preprints202503.0658.v1



(a) Shallow mixed layers at 1°S and 30°W. **(b)** Deep mixed layers at 49°N and 29°W.

Figure 10. Depth-time interpolation from BSFDA$^{\text{Fast}}$ vs. actual ARGO global oceanic measurements at two sites focusing on mixed layer behaviors. Measurements are represented by circles, with filling color indicating temperature. Circle sizes show distance in geodesic space from the central point.

To our knowledge, this is the first time the ARGO dataset has been modeled in a full 4-dimensional principal component model, with the correct domain topology. We incorporate the entire period of 27 years rather than shorter spans (e.g., 2004-2008 or 2007-2016) [78,80,81]. Instead of segmenting the dataset into localized spatiotemporal windows, we process the entire 4D domain (latitude, longitude, depth, and intra-annual time) in a single holistic framework. Previous studies were typically tailored for ARGO datasets and handled each depth, month, or spatial region separately, restricting correlation estimates to limited windows (e.g., 1000 km and three months) while excluding data with large offsets [78,81]. In addition, they require repeated on-demand model fitting that can hinder scalability. Our kernel-based framework, by contrast, is broadly applicable to general functional data, only requiring kernel definitions for the domain. Although global dimension-reduction inevitably introduces some residual noise, the kernel-based design is extensible to finer spacing or multiple length scales if higher precision is needed. Furthermore, inference with our model is simply the evaluation of the active 163 active basis functions weighted by the 16 principal components. Interpolation over a 300×300 grid only takes about two seconds. By contrast, previous methods with Gaussian process regression require a weighted sum of all the measured data within a certain window. The parametric representations also facilitate straightforward derivative and integration calculations, which are essential for investigating ocean temperature stratification and heat content [78]. In summary, the Argo dataset provides an ideal testbed for our method, as it captures the dynamic behavior of high-dimensional geospatial data in a continuous framework. A more comprehensive study of ARGO is beyond this paper's scope. Nonetheless, the results here confirm the clear advantages of the proposed method for large-scale, high-dimensional functional data.

6. Discussion

This paper proposed BSFDA, a novel framework for functional data analysis with irregular sampling, integrating model selection and scalability in one unique, coherent, and effective algorithm. Our extensive empirical studies, including both simulations and real world applications, show that BSFDA offer superior covariance estimation accuracy with remarkable efficiency.

In terms of accuracy, our method excels in model selection, consistently achieving top-tier performance. The accuracy of the covariance operator estimation also rivals that of the best existing methodologies in the field. This shows that our approach can not only handle large and complex datasets, but also ensures high accuracy and precision in the results it produces. Our method's superiority compared to existing techniques is expected owing to the inherent iterative nature of data smoothing and covariance estimation in our approach.

 doi:10.20944/preprints202503.0658.v1

In terms of scalability, our method demonstrates a linear growth of time complexity with the size of the dataset, and impressively, the computations are executed in a small, $K^{(a)}$-dimensional subspace. This ensures that as the datasets grow larger and more complex, the performance of our model remains robust and efficient. Additionally, we introduced a faster variant, BSFDA$^{\text{Fast}}$, which performs similarly to BSFDA on medium and dense datasets with significantly reduced computational cost. This leap in efficiency enabled a full 4-dimensional functional modeling, for the first time, of the large scale oceanic temperature dataset across 27 years (ARGO) [71]. Although BSFDA$^{\text{Fast}}$ can underestimate signal strength under very sparse sampling, the vanilla BSFDA effectively complements and alleviates this issue.

Looking ahead, it would be interesting to explore how extensions of regular PCA, such as simplified PCA and robust PCA [42], can be integrated within our proposed framework. These extensions will enhance the flexibility and robustness of our method, further improving its adaptability to various data conditions. In addition, we see potential in examining the extensions of functional PCA, such as time warping, dynamics, and manifold learning [1]. In particular, shape analysis emerges as a direct application of time warping. Such extensions would push the boundaries of what our proposed method could achieve, potentially enabling it to handle an even wider array of data structures and complexities.

In conclusion, our research findings affirm the proposed framework's effectiveness and adaptability in advanced functional data analysis. Nonetheless, the method's potential remains broad, and future work promises to widen its scope and refine its performance. By unifying sparse Bayesian learning, kernel-based expansions, and efficient variational inference, BSFDA offers a powerful foundation for large-scale, high-dimensional FDA challenges.

7. System of Notation

Table 7 summarizes the notation used in Section 1.2 and 3.2.3, providing a reference for the derivations. All vectors in the table are represented as row vectors.

Table 7. Symbol definitions in formulation.

Symbol	Meaning
y_i	The i-th sample function
$x \in R^M$	One M-dimension index
M	Dimension of the index set
K	Number of all basis functions
J	Number of all components
P	Number of sample functions
N_i	Number of measurements of the i-th sample function
$X_i \in R^{Ni \times M}$	Index set of the i-th sample function
$Y_i \in R^{Ni}$	Measurement of the i-th sample function
$Z_i \in R^J$	Component scores of the i-th sample function
$\bar{Z} \in R^K$	Coefficients of basis functions in the mean function
$E_i \in R^{Ni}$	Measurement errors of the i-th sample function
$W \in R^{J \times K}$	Weighing matrix of basis functions in the eigenfunctions
$W_{j\cdot} \in R^K, W_{\cdot k}^T \in R^J$	The j-th row and k-th column of W
$\mathcal{K}$	The kernel function
α_j	The scale parameter of $W_{j\cdot}$ (j-th component)
β_k	The scale parameter of $W_{\cdot k}$ (k-th basis function)
σ	The standard deviation of measurement errors
η	The communal scale parameter of $\bar{Z}$
$\{\phi_k : R^M \to R\}_{k=1}^K$	The union of all the centered kernel functions
$\Phi_{ikj} = \phi_k(X_{ij\cdot}) \in R$	Value of centered kernel function ϕ_k at $X_{ij\cdot}$
$\theta_i \in R^K$	Coefficients of the i-th sample function
$\zeta_i \in R^K$	Coefficient noise of the i-th sample function
ς_k	The scale parameter of k-th coefficient noise

 doi:10.20944/preprints202503.0658.v1

Table 8 summarizes the notation used in Section 2.2.

Table 8. Notation used in formulating the optimization.

Symbol	Meaning
Θ	All the latent variables.
$\mathcal{Q}_{\cdot}$	The surrogate posterior distribution of variable $\cdot$
$\mathcal{Q}_{/\cdot}$	The joint surrogate posterior distribution of all variables except $\cdot$
$\mu_{\cdot}, \Sigma_{\cdot}$	The mean and covariance of $\cdot$ in $\mathcal{Q}$, e.g. $\mu_{\mathrm{vec}(W)} \in R^{JK}, \Sigma_{\mathrm{vec}(W)} \in R^{JK \times JK}$
$a_{\cdot}, b_{\cdot}$	The shape and rate parameters of $\mathcal{Q}_{\cdot}$, e.g. a_{β_k}, b_{β_k}
$\mathbb{E}_{\mathcal{Q}}[\cdot]$	The expectation of variable $\cdot$ over density $\mathcal{Q}$
$\mathcal{L}$	The lower bound of surrogate posterior $\mathcal{Q}$ with K basis functions
Ψ_i	Gram matrix of the kernel functions for the i-th sample function, $\Phi_i \Phi_i^T$
$K^{(\mathrm{a})}, K^{(\mathrm{e})}$	Number of active/effective basis functions
$J^{(\mathrm{a})}, J^{(\mathrm{e})}$	Number of active/effective components
$\mathcal{P}_i$	Log likelihood of Y_i in multi-sample relevance vector machine
$\mathcal{C}_i$	Covariance of Y_i in multi-sample relevance vector machine
$\mathcal{S}_i$	Posterior covariance of Z_i in multi-sample relevance vector machine
$\mathcal{P}_{Z_i}$	Log likelihood of (Y_i, Z_i) in multi-sample relevance vector machine
$\epsilon \to 0$	The infinitesimal number
$\tau_{\cdot}$	Threshold/tolerance of $\cdot$

8. Variational Update Formulae

As defined in Section 1.2, we consider the following priors and conditional distributions:

$$\Pr[Y|Z, W, \bar{Z}, \sigma] = \prod_i \mathcal{N}\left(Y_i|(Z_i W + \bar{Z})\Phi_i, \sigma^2 I\right) \tag{45}$$

$$\Pr[Z] = \prod_i \mathcal{N}(Z_i|0, I) \tag{46}$$

$$\Pr[W|\alpha, \beta] = \prod_{j,k} \mathcal{N}(W_{jk}|0, \alpha_j^{-1}\beta_k^{-1}) \tag{47}$$

$$\Pr[\bar{Z}] = \prod_k \mathcal{N}(\bar{Z}_k|0, \eta^{-1}\beta_k^{-1}) \tag{48}$$

$$\Pr[\sigma]\Pr[\alpha]\Pr[\beta]\Pr[\eta] = \Gamma(\sigma^{-2}|a_0, b_0)\prod_{j=1}^{J}\Gamma(\alpha_j|a_0, b_0)\prod_{k=1}^{K}\Gamma(\beta_k|a_0, b_0)\Gamma(\eta|a_0, b_0) \tag{49}$$

For brevity, the joint posterior is shown with the vague Gamma prior parameters a_0, b_0, and the observation index X omitted:

$$\begin{aligned}
&\Pr[Z, W, \bar{Z}, \sigma, \alpha, \beta, \eta|X, Y, a_0, b_0] = \Pr[Z, W, \bar{Z}, \sigma, \alpha, \beta, \eta|Y] \\
&= \Pr[Z, W, \bar{Z}, \sigma, \alpha, \beta, \eta, Y](\Pr[Y])^{-1} \\
&\propto \Pr[Z, W, \bar{Z}, \sigma, \alpha, \beta, \eta, Y] \\
&= \Pr[Y|Z, W, \bar{Z}, \sigma]\Pr[Z]\Pr[W|\alpha, \beta]\Pr[\bar{Z}|\eta, \beta]\Pr[\sigma]\Pr[\alpha]\Pr[\beta]\Pr[\eta]
\end{aligned} \tag{50}$$

Derivation of Equations (14) and (15):

 doi:10.20944/preprints202503.0658.v1

According to Equation (13) and the posterior in Equation (50), the update formulae for the surrogate distribution $\mathcal{Q}_{\alpha_j}$ is:

$$\begin{aligned}
\mathcal{Q}_{\alpha_j} &\leftarrow \frac{\exp\left(\mathbb{E}_{\mathcal{Q}_{/\alpha_j}}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right)}{\int \exp\left(\mathbb{E}_{\mathcal{Q}_{/\alpha_j}}[\ln(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y])]\right) d\alpha_j} \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\alpha_j}}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right) \propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\alpha_j}}[\ln\left(\Pr[W|\alpha,\beta]\Pr[\alpha]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\alpha_j}}\left[-\frac{1}{2}\sum_{k=1}^{K}\left(-\ln\left(\alpha_j\right)+W_{jk}^2\alpha_j\beta_k\right)+\left(\left(a_0-1\right)\ln\alpha_j-b_0\alpha_j\right)\right]\right) \\
&\propto \exp\left(\left(\frac{K}{2}+a_0-1\right)\ln\left(\alpha_j\right)-\alpha_j\left(\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{Q}_{/\alpha_j}}\left[\left(W_{jk}^2\beta_k\right)\right]+b_0\right)\right)
\end{aligned} \tag{51}$$

where we have omitted terms that α_j is conditionally independent of. By definition

$$\begin{aligned}
\mathcal{Q}_{\alpha_j} &= \exp\Big(\ln(\Gamma(\alpha_j|a_{\alpha_j},b_{\alpha_j}))\Big) = \exp\left(\ln\left(\frac{b_{\alpha_j}^{a_{\alpha_j}}}{\Gamma(a_{\alpha_j})}\alpha_j^{a_{\alpha_j}-1}\exp\left(-b_{\alpha_j}\alpha_j\right)\right)\right) \\
&\propto \exp\Big(\left(a_{\alpha_j}-1\right)\ln\alpha_j-b_{\alpha_j}\alpha_j\Big)
\end{aligned} \tag{52}$$

By equating Equations (51) and (52), the updates for $\mathcal{Q}_{\alpha_j}$ are

$$a_{\alpha_j} \leftarrow \frac{K}{2}+a_0 \tag{53}$$

$$b_{\alpha_j} \leftarrow \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{Q}_{/\alpha_j}}\left[\left(W_{jk}^2\beta_k\right)\right]+b_0 \tag{54}$$

Derivation of Equations (16) and (17):
According to Equation (13) and the posterior Equation (50), the update formulae for $\mathcal{Q}_{\eta}$ is:

$$\begin{aligned}
\mathcal{Q}_{\eta} &\leftarrow \frac{\exp\left(\mathbb{E}_{\mathcal{Q}_{/\eta}}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right)}{\int \exp\Big(\mathbb{E}_{\mathcal{Q}_{/\eta}}[\ln(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y])]\Big) d\eta} \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}/\eta}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right) \propto \exp\left(\mathbb{E}_{\mathcal{Q}/\eta}[\ln\left(\Pr[\bar{Z}|\eta,\beta]\Pr[\eta]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}/\eta}\left[-\frac{1}{2}\sum_{k=1}^{K}\left(-\ln\left(\eta\right)+\bar{Z}_k^2\eta\beta_k\right)+\left(\left(a_0-1\right)\ln\eta-b_0\eta\right)\right]\right) \\
&\propto \exp\left(\left(\frac{K}{2}+a_0-1\right)\ln\left(\eta\right)-\eta\left(\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{Q}/\eta}\left[\left(\bar{Z}_k^2\beta_k\right)\right]+b_0\right)\right)
\end{aligned} \tag{55}$$

where we have omitted terms η is conditionally independent of. By definition

$$\begin{aligned}
\mathcal{Q}_{\eta} &= \exp\Big(\ln(\Gamma(\eta|a_{\eta},b_{\eta}))\Big) = \exp\left(\ln\left(\frac{b_{\eta}^{a_{\eta}}}{\Gamma(a_{\eta})}\eta^{a_{\eta}-1}\exp\left(-b_{\eta}\eta\right)\right)\right) \\
&\propto \exp\Big(\left(a_{\eta}-1\right)\ln\eta-b_{\eta}\eta\Big)
\end{aligned} \tag{56}$$

 doi:10.20944/preprints202503.0658.v1

By equating Equations (55) and (56), the updates for $\mathcal{Q}_\eta$ are

$$a_\eta \leftarrow \frac{K}{2} + a_0 \tag{57}$$

$$b_\eta \leftarrow \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}_{\mathcal{Q}/\eta}\left[\left(\bar{Z}_k^2 \beta_k\right)\right] + b_0 \tag{58}$$

Derivation of Equations (18), (19):
According to Equation (13) and the posterior Equation (50), the update formulae for $\mathcal{Q}_{\beta_k}$ is:

$$\begin{aligned}
\mathcal{Q}_{\beta_k} &\leftarrow \frac{\exp\left(\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[\ln\left(\Pr[Z, W, \bar{Z}, \sigma, \alpha, \beta, \eta, Y]\right)\right]\right)}{\int \exp\left(\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[\ln(\Pr[Z, W, \bar{Z}, \sigma, \alpha, \beta, \eta, Y])\right]\right) d\beta_k} \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[\ln\left(\Pr[Z, W, \bar{Z}, \sigma, \alpha, \beta, \eta, Y]\right)\right]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[\ln\left(\Pr[W_{\cdot k}|\alpha, \beta_k]\Pr[\bar{Z}_k|\eta, \beta_k]\Pr[\beta]\right)\right]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[-\frac{1}{2}\sum_{j=1}^{J}\left(-\ln(\alpha_j\beta_k) + W_{jk}^2\alpha_j\beta_k\right)\right.\right. \\
&\quad \left.\left. -\frac{1}{2}\left(-\ln(\eta\beta_k) + \bar{Z}_k^2\eta\beta_k\right) + (a_0 - 1)\ln\beta_k - b_0\beta_k\right]\right) \\
&\propto \exp\left(\left(\frac{J+1}{2} + a_0 - 1\right)\ln(\beta_k) - \right. \\
&\left. \beta_k\left(\frac{1}{2}\left(\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[\left(\bar{Z}_k^2\eta\right)\right] + \sum_{j=1}^{J}\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[\left(W_{jk}^2\alpha_j\right)\right]\right) + b_0\right)\right)
\end{aligned} \tag{59}$$

where we have omitted terms that β_k in conditionally independent of. By definition

$$\begin{aligned}
\mathcal{Q}_{\beta_k} &= \exp\left(\ln(\Gamma(\beta_k|a_{\beta_k}, b_{\beta_k}))\right) = \exp\left(\ln\left(\frac{b_{\beta_k}^{a_{\beta_k}}}{\Gamma(a_{\beta_k})}\eta^{a_{\beta_k}-1}\exp\left(-b_{\beta_k}\eta\right)\right)\right) \\
&\propto \exp\left((a_{\beta_k} - 1)\ln\eta - b_{\beta_k}\eta\right)
\end{aligned} \tag{60}$$

By equating Equations (59) and (60), the updates for $\mathcal{Q}_\eta$ are

$$a_{\beta_k} \leftarrow \frac{J+1}{2} + a_0 \tag{61}$$

$$b_{\beta_k} \leftarrow \frac{1}{2}\left(\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[\left(\bar{Z}_k^2\eta\right)\right] + \sum_{j=1}^{J}\mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[\left(W_{jk}^2\alpha_j\right)\right]\right) + b_0 \tag{62}$$

Derivation of Equations (20), (21):

 doi:10.20944/preprints202503.0658.v1

According to Equations (13) and the posterior Equation (50), the update formulae for $\mathcal{Q}_{\bar{Z}}$ is:

$$\begin{aligned}
\mathcal{Q}_{\bar{Z}} &\leftarrow \frac{\exp\left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right)}{\int \exp\Big(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[\ln(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y])]\Big)d\bar{Z}} \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[\ln\left(\Pr[Y|Z,W,\bar{Z},\sigma]\Pr[\bar{Z}|\eta,\beta]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[-\frac{1}{2}\sum_{i=1}^{P}\Big(N_i\ln(2\pi\sigma^2)+\sigma^{-2}||Y-(Z_iW+\bar{Z})\Phi_i||_2^2\Big)-\right.\right. \\
&\left.\left.\frac{1}{2}\sum_{k=1}^{K}\Big(-\ln\left(2\pi\eta\beta_k\right)+\bar{Z}_k^2\eta\beta_k\Big)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\bar{Z}\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\sum_{i=1}^{P}\Psi_i+\eta\,\mathrm{diag}(\beta)\right]\bar{Z}^T-\right.\right. \\
&\left.\left.2\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\right]\sum_{i=1}^{P}(Y-\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[Z_iW]\Phi_i)\Phi_i^T\bar{Z}^T\right)\right)
\end{aligned} \tag{63}$$

where we have omitted terms that $\bar{Z}$ is conditionally independent of. By definition

$$\begin{aligned}
\mathcal{Q}_{\bar{Z}} &= \exp(\ln(\mathcal{N}(\bar{Z}|\mu_{\bar{Z}},\Sigma_{\bar{Z}}))) = \exp\left(-\frac{1}{2}\Big(\ln|2\pi\Sigma_{\bar{Z}}|+(\bar{Z}-\mu_{\bar{Z}})\Sigma_{\bar{Z}}{}^{-1}(\bar{Z}-\mu_{\bar{Z}})^T\Big)\right) \\
&\propto \exp\left(-\frac{1}{2}\Big(\bar{Z}\Sigma_{\bar{Z}}{}^{-1}\bar{Z}^T-2\mu_{\bar{Z}}\Sigma_{\bar{Z}}{}^{-1}\bar{Z}^T\Big)\right)
\end{aligned} \tag{64}$$

By equating Equations (63) and (64), the updates for $\mathcal{Q}_{\bar{Z}}$ are

$$\Sigma_{\bar{Z}} \leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\sum_{i=1}^{P}(\Psi_i)+\eta\,\mathrm{diag}(\beta)\right]\right)^{-1} \tag{65}$$

$$\mu_{\bar{Z}} \leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\right]\sum_{i=1}^{P}(Y-\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[Z_iW]\Phi_i)\Phi_i^T\right)\Sigma_{\bar{Z}} \tag{66}$$

Derivation of Equations (22), (23):

 doi:10.20944/preprints202503.0658.v1

According to Equation (13) and the posterior Equation (50), the update formulae for $\mathcal{Q}_W$ is:

$$
\begin{aligned}
\mathcal{Q}_W &\leftarrow \frac{\exp\left(\mathbb{E}_{\mathcal{Q}_{/W}}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right)}{\int \exp\Big(\mathbb{E}_{\mathcal{Q}_{/W}}[\ln(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y])]\Big)dW} \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/W}}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/W}}[\ln\left(\Pr[Y|Z,W,\bar{Z},\sigma]\Pr[W|\alpha,\beta]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/W}}\left[-\frac{1}{2}\sum_{i=1}^{P}\Big(N_i\ln(2\pi\sigma^2)+\sigma^{-2}||Y-(Z_iW+\bar{Z})\Phi_i||_2^2\Big)\right]\right) \\
&\exp\left(\mathbb{E}_{\mathcal{Q}_{/W}}\left[-\frac{1}{2}\Big(\ln|2\pi(\mathrm{diag}(\beta)\otimes\mathrm{diag}(\alpha))^{-1}|+\right.\right. \\
&\left.\left.\mathrm{vec}(W)^T(\mathrm{diag}(\beta)\otimes\mathrm{diag}(\alpha))\,\mathrm{vec}(W)\Big)\right]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}_{/W}}\left[-\frac{1}{2}\Big(\sigma^{-2}\sum_{i=1}^{P}\Big(-2Y_i\Phi_i^TW^TZ_i^T+2Z_iW\Psi_i\bar{Z}^T+Z_iW\Psi_iW^TZ_i^T\Big)\Big)\right]\right) \\
&\exp\left(\mathbb{E}_{\mathcal{Q}_{/W}}\left[-\frac{1}{2}\Big(\mathrm{vec}(W)^T(\mathrm{diag}(\beta)\otimes\mathrm{diag}(\alpha))\,\mathrm{vec}(W)\Big)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\mathbb{E}_{\mathcal{Q}_{/W}}\left[-2\sigma^{-2}\sum_{i=1}^{P}\mathrm{vec}\Big(\Big(\Phi_i(\Phi_i^T\bar{Z}^T-Y_i^T)Z_i\Big)^T\Big)^T\right]\mathrm{vec}(W)\right) \\
&\exp\left(-\frac{1}{2}\mathrm{vec}(W)^T\mathbb{E}_{\mathcal{Q}_{/W}}\left[\sigma^{-2}\sum_{i=1}^{P}\Big(\Psi\otimes(Z_i^TZ_i)\Big)+(\mathrm{diag}(\beta)\otimes\mathrm{diag}(\alpha))\right]\mathrm{vec}(W)\right) \qquad (67)
\end{aligned}
$$

where we have omitted terms that W is conditionally independent of. By definition

$$
\begin{aligned}
\mathcal{Q}_W &= \exp\Big(\ln(\mathcal{N}(\mathrm{vec}(W)|\mu_{\mathrm{vec}(W)},\Sigma_{\mathrm{vec}(W)}))\Big) \\
&= \exp\left(-\frac{1}{2}\Big(\ln|2\pi\Sigma_{\mathrm{vec}(W)}|+\right. \\
&\left.(\mathrm{vec}(W)^T-\mu_{\mathrm{vec}(W)})\Sigma_{\mathrm{vec}(W)}{}^{-1}(\mathrm{vec}(W)^T-\mu_{\mathrm{vec}(W)})^T\Big)\right) \\
&\propto \exp\Big(-\frac{1}{2}\Big(\mathrm{vec}(W)^T\Sigma_{\mathrm{vec}(W)}{}^{-1}\,\mathrm{vec}(W)-2\mu_{\mathrm{vec}(W)}\Sigma_{\mathrm{vec}(W)}{}^{-1}\,\mathrm{vec}(W)\Big)\Big) \qquad (68)
\end{aligned}
$$

By equating Equations (67) and (68), the updates for $\mathcal{Q}_W$ are

$$
\begin{aligned}
\Sigma_{\mathrm{vec}(W)} &\leftarrow \mathbb{E}_{\mathcal{Q}_{/W}}\left[\sigma^{-2}\sum_{i=1}^{P}\Big(\Psi\otimes(Z_i^TZ_i)\Big)+(\mathrm{diag}(\beta)\otimes\mathrm{diag}(\alpha))\right]^{-1} \\
\mu_{\mathrm{vec}(W)} &\leftarrow \mathbb{E}_{\mathcal{Q}_{/W}}\left[-\sigma^{-2}\sum_{i=1}^{P}\mathrm{vec}\Big(\Big(\Phi_i(\Phi_i^T\bar{Z}^T-Y_i^T)Z_i\Big)^T\Big)^T\right]\Sigma_{\mathrm{vec}(W)} \qquad (69)
\end{aligned}
$$

Derivation of Equations (24), (25) and (26):

 doi:10.20944/preprints202503.0658.v1

According to Equation (13) and the posterior Equation (50), the update formulae for $\mathcal{Q}_{Z_i}$ is:

$$\begin{aligned}
\mathcal{Q}_{Z_i} &\leftarrow \frac{\exp\left(\mathbb{E}_{\mathcal{Q}/Z_i}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right)}{\int \exp\Big(\mathbb{E}_{\mathcal{Q}/Z_i}[\ln(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y])]\Big)dZ_i} \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}/Z_i}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}/Z_i}[\ln\left(\Pr[Y_i|Z_i,W,\bar{Z},\sigma]\Pr[Z_i]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}/Z_i}\left[-\frac{1}{2}\Big(N_i\ln(2\pi\sigma^2)+\sigma^{-2}||Y-(Z_iW+\bar{Z})\Phi_i||_2^2+J\ln(2\pi)+Z_iZ_i^T\Big)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left(Z_i\mathbb{E}_{\mathcal{Q}/Z_i}\left[\sigma^{-2}W\Psi_iW^T+I\right]Z_i^T-2\mathbb{E}_{\mathcal{Q}/Z_i}\left[\sigma^{-2}(Y_i-Z_i\Phi_i)\Phi_i^TW^T\right]Z_i^T\right)\right) \qquad (70)
\end{aligned}$$

where we have omitted terms that Z_i is conditionally independent of. By definition

$$\begin{aligned}
\mathcal{Q}_{Z_i} &= \exp\big(\ln(\mathcal{N}(Z_i|\mu_{Z_i},\Sigma_{Z_i}))\big) \\
&= \exp\left(-\frac{1}{2}\Big(\ln|2\pi\Sigma_{Z_i}|+(Z_i-\mu_{Z_i})\Sigma_{Z_i}{}^{-1}(Z_i-\mu_{Z_i})^T\Big)\right) \\
&\propto \exp\left(-\frac{1}{2}\Big(Z_i\Sigma_{Z_i}{}^{-1}Z_i^T-2\mu_{Z_i}\Sigma_{Z_i}{}^{-1}Z_i^T\Big)\right) \qquad (71)
\end{aligned}$$

By equating Equations (70) and (71), the updates for $\mathcal{Q}_{\bar{Z}}$ are

$$\Sigma_{Z_i} \leftarrow \left(\mathbb{E}_{\mathcal{Q}/Z_i}[\sigma^{-2}W\Psi_iW^T+I]\right)^{-1} \qquad (72)$$

$$\mu_{Z_i} \leftarrow \mathbb{E}_{\mathcal{Q}/Z_i}[\sigma^{-2}(Y_i-\bar{Z}\Phi_i)\Phi_i^TW^T]\Sigma_{Z_i} \qquad (73)$$

Derivation of Equations (27), (28) and (29):
According to Equation (13) and the posterior Equation (50), the update formulae for $\mathcal{Q}_\sigma$ is:

$$\begin{aligned}
\mathcal{Q}_\sigma &\leftarrow \frac{\exp\left(\mathbb{E}_{\mathcal{Q}/\sigma}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right)}{\int \exp\Big(\mathbb{E}_{\mathcal{Q}/\sigma}[\ln(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y])]\Big)d\sigma} \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}/\sigma}[\ln\left(\Pr[Z,W,\bar{Z},\sigma,\alpha,\beta,\eta,Y]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}/\sigma}[\ln\left(\Pr[Y_i|Z_i,W,\bar{Z},\sigma]\Pr[\sigma]\right)]\right) \\
&\propto \exp\left(\mathbb{E}_{\mathcal{Q}/\sigma}\left[-\frac{1}{2}\sum_{i=1}^{P}\Big(N_i\ln(2\pi\sigma^2)+\sigma^{-2}||Y-(Z_iW+\bar{Z})\Phi_i||_2^2\Big)\right]\right) \\
&\exp\left(\mathbb{E}_{\mathcal{Q}/\sigma}\left[\Big((a_0-1)\ln\sigma^{-2}-b_0\sigma^{-2}\Big)\right]\right) \\
&\propto \exp\left(\left(a_0+\frac{1}{2}\sum_i N_i-1\right)\ln(\sigma^{-2})-\right. \\
&\left.\sigma^{-2}\left(b_0+\frac{1}{2}\mathbb{E}_{\mathcal{Q}/\sigma}[\sum_i||Y_i-(Z_iW+\bar{Z})\Phi_i||_2^2]\right)\right) \qquad (74)
\end{aligned}$$

where we have omitted terms that σ is conditionally independent of. By definition

$$\begin{aligned}
\mathcal{Q}_\sigma &= \exp\Big(\ln(\Gamma(\sigma^{-2}|a_\sigma,b_\sigma))\Big) = \exp\left(\ln\left(\frac{{b_\sigma}^{a_\sigma}}{\Gamma(a_\sigma)}(\sigma^{-2})^{a_\sigma-1}\exp\left(-b_\sigma\sigma^{-2}\right)\right)\right) \\
&\propto \exp\Big((a_\sigma-1)\ln\sigma^{-2}-b_\sigma\sigma^{-2}\Big) \qquad (75)
\end{aligned}$$

 doi:10.20944/preprints202503.0658.v1

By equating Equations (74) and (75), the updates for $\mathcal{Q}_\sigma$ are

$$a_\sigma \leftarrow a_0 + \frac{1}{2}\sum_i N_i \tag{76}$$

$$b_\sigma \leftarrow b_0 + \frac{1}{2}\mathbb{E}_{\mathcal{Q}_{/\sigma}}\left[\sum_i ||Y_i - (Z_i W + \bar{Z})\Phi_i||_2^2\right] \tag{77}$$

9. Scalable Update for BSFDA

9.1. Implicit Factorization

We initialize the inactive precision parameters as:

$$\mathbb{E}_{\mathcal{Q}_{\alpha_j}}[\alpha_j] = \epsilon^{-1}, \forall j > J^{(\mathrm{a})} \tag{78}$$

$$\mathbb{E}_{\mathcal{Q}_{\beta_k}}[\beta_k] = \epsilon^{-1}, \forall k > K^{(\mathrm{a})} \tag{79}$$

Under these settings and subsequent variational updates (using Equations (78) and (79)), in the limit as $\epsilon \to 0$, the surrogate distributions satisfy:

$$\mu_{Z_{iB}} = 0,\ \Sigma_{Z_i B} = \epsilon I^{(J-J^{(\mathrm{a})})},\ \Sigma_{Z_i[A,B]} = \left(\Sigma_{Z_i[B,A]}\right)^T = 0 \tag{80}$$

$$\mu_{\bar{Z}B} = 0,\ \Sigma_{\bar{Z}[B,B]} = \epsilon I^{(K-K^{(\mathrm{a})})},\ \Sigma_{\bar{Z}[A,B]} = \Sigma_{\bar{Z}[B,A]}^T = 0 \tag{81}$$

$$\begin{aligned}
&\mu_{\mathrm{vec}(W)_B} = 0,\ \mu_{\mathrm{vec}(W)_C} = 0,\ \mu_{\mathrm{vec}(W)_D} = 0,\ \Sigma_{\mathrm{vec}(W)[B,B]} = \epsilon I^{(K^{(\mathrm{a})}J - K^{(\mathrm{a})}J^{(\mathrm{a})})},\\
&\Sigma_{\mathrm{vec}(W)[C,C]} = \epsilon I^{(J^{(\mathrm{a})}K - J^{(\mathrm{a})}K^{(\mathrm{a})})},\ \Sigma_{\mathrm{vec}(W)[D,D]} = \epsilon I^{(JK + J^{(\mathrm{a})}K^{(\mathrm{a})} - JK^{(\mathrm{a})} - J^{(\mathrm{a})}K)},\\
&\Sigma_{\mathrm{vec}(W)[x,y]} = 0, \forall (x,y) \notin \{(A,A),(B,B),(C,C),(D,D)\}
\end{aligned} \tag{82}$$

For convenience, we initialize $\mathcal{Q}$ with the above properties.

Lemma 1. *If $\mathcal{Q}_{\alpha_j}[\alpha_j] = \epsilon, \forall j \geq J^{(a)}$ and $\mathcal{Q}_{\beta_k}[\beta_k] = \epsilon, \forall k \geq K^{(a)}$, then the variational distribution over W factorizes as $\mathcal{Q}_W = \mathcal{Q}_{W_A}\mathcal{Q}_{W_B}\mathcal{Q}_{W_C}\mathcal{Q}_{W_D}$ in the limit as $\epsilon \to 0$.*

Proof. We express the distribution as

$$\begin{aligned}
\mathcal{Q}_W &= \mathcal{N}(\mathrm{vec}(W)|\mu_{\mathrm{vec}(W)}, \Sigma_{\mathrm{vec}(W)})\\
&= \exp\left(-\frac{1}{2}\left(\ln|2\pi\Sigma_{\mathrm{vec}(W)}| + \mu_{\mathrm{vec}(W)}{\Sigma_{\mathrm{vec}(W)}}^{-1}{\mu_{\mathrm{vec}(W)}}^T\right)\right),
\end{aligned}$$

The factorization holds if the off-diagonal block matrices in $\Sigma_{\mathrm{vec}(W)}$, e.g. $\Sigma_{[W_A,W_B]}$, are all zero, i.e., the blocks are mutually independent. Initially, this is ensured by the definition for the initial status in Equation (82). Thus, we only need to show the statement remains true after $\mathcal{Q}_W$ is updated, i.e., after Equation (22) is applied with the inactive scale parameters $\mathcal{Q}_{\alpha_j}[\alpha_j]$ and $\mathcal{Q}_{\beta_k}[\beta_k]$ fixed at ϵ. First we regard $\Sigma_{[W_{ABC}]}$, i.e., the covariance of the union of W_A, W_B, W_C after vectorization, as one block. By the block matrix inversion formula, we get $\Sigma_{[W_{ABC},W_D]} \propto \epsilon^2 \to 0$ and consequently $\mathcal{Q}_W = \mathcal{Q}_{W_{ABC}}\mathcal{Q}_{W_D}$. Next, apply block matrix inversion formula to $\Sigma_{[W_{ABC}]}$ in Equation (22) and we get $(\Sigma_{[W_B,W_A]}, \Sigma_{[W_B,W_C]}, \Sigma_{[W_C,W_A]}, \Sigma_{[W_C,W_B]}) \propto \epsilon \to 0$, yielding the desired factorization. □

Lemma 2. *If $\mathcal{Q}_{\beta_k}[\beta_k] = \epsilon, \forall k \geq K^{(a)}$, then the implicit factorization $\mathcal{Q}_{\bar{Z}} = \mathcal{Q}_{\bar{Z}_A}\mathcal{Q}_{\bar{Z}_B}$ holds in the limit as $\epsilon \to 0$.*

 doi:10.20944/preprints202503.0658.v1

Proof. It is similar to the proof for Lemma 1. Because $\mathcal{Q}_{\bar{Z}} = \mathcal{N}(\mu_{\bar{Z}}, \Sigma_{\bar{Z}})$, we only need the off-diagonal block is zero, i.e., $\Sigma_{\bar{Z}[A,B]} = 0$. Initially, it is ensured by definition for the initial status in Equation (81). $\mathcal{Q}_{\bar{Z}}$ is updated by Equation (20), Applying block matrix inversion formula with the inactive $\mathcal{Q}_{\beta_k}[\beta_k]$, we get $\Sigma_{\bar{Z}[A,B]} \propto \epsilon \to 0$, establishing the factorization. □

Lemma 3. *If $j \geq J^{(a)}$ or $k \geq K^{(a)}$, then $\mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_{kj'}W_{jk'}] \propto \mathcal{O}(\epsilon), \forall j' = 1:J, k' = 1:K$ in the limit as $\epsilon \to 0$.*

Proof. For initial status, apparently the largest $\mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_{kj'}W_{jk'}]$ is $\mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_{kj'}^2] = \epsilon$. Because either $\mathcal{Q}_{\alpha_j}[\alpha_j] = \epsilon$ or $\mathcal{Q}_{\beta_k}[\beta_k] = \epsilon$, after updates from Equations (22) and (23) are applied, $\mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_{kj'}W_{jk'}] = \Sigma_{[W_{kj'},W_{jk'}]} + \mu_{\mathrm{vec}(W)_{kj'}}\mu_{\mathrm{vec}(W)_{jk'}} \propto \mathcal{O}(\epsilon)$ by Woodbury matrix identity. □

Lemma 4. *If $\mathbb{E}_{\mathcal{Q}_{\alpha_j}}[\alpha_j] = \frac{a_{\alpha_j}}{b_{\alpha_j}} = \epsilon^{-1}, \forall j \geq J^{(a)}$, then the implicit factorization $\mathcal{Q}_{Z_i} = \mathcal{Q}_{Z_{iA}}\mathcal{Q}_{Z_{iB}}$ holds in the limit as $\epsilon \to 0$.*

Proof. It is similar to the proof for Lemma 1. Because $\mathcal{Q}_{Z_i} = \mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i})$, only $\Sigma_{Z_i[A,B]} = 0$ is needed. Initially, it is ensured by definition for the initial status Equation (80). $\mathcal{Q}_Z$ is updated by Equations (24) and (25). In Equation (24), when $j \geq J^{(a)}$ or $k \geq K^{(a)}$, $C_{ijk} = \mathrm{Tr}(\mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_{k\cdot}^T W_{j\cdot}]\Psi_i) = \sum_{(j',k')}\left(\mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_{kj'}W_{jk'}](\Psi_i)_{k'j'}\right) \propto \mathcal{O}(\epsilon) \to 0$ applying Lemma 3. Applying block matrix inversion formula to Equation (25), $\Sigma_{Z_i[AB]} \propto \mathcal{O}(\epsilon) \to 0$, , thus proving the implicit factorization □

9.2. Scale Parameters

Here we state the theorems that justify we can use updating rules for $\mathcal{Q}^{(a)}{}_{\alpha_j}$ based on $\mathcal{L}^{(a)}$ to update $\mathcal{Q}_{\alpha_j}$ (and similarly, $\mathcal{Q}^{(a)}{}_{\beta_k}$ for $\mathcal{Q}_{\beta_k}$, $\mathcal{Q}^{(a)}{}_{\eta}$ for $\mathcal{Q}_\eta$) and it does maximize $\mathcal{L}$ ultimately.

Lemma 5. *$\forall W_{jk} \in W_B \cup W_C$, i.e., either $(j > J^{(a)})$ or $(k > K^{(a)})$, after updating $\mathcal{Q}_{W_B}$ and $\mathcal{Q}_{W_C}$ by Equations (22) and (23), $\mathbb{E}_{\mathcal{Q}}[W_{jk}^2] = \frac{b_{\alpha_j}b_{\beta_k}}{a_{\alpha_j}a_{\beta_k}}$.*

Proof. According to Equations (78) and (79), if $(j > J^{(a)})$ or $(k > K^{(a)})$, either $\mathbb{E}_{\mathcal{Q}}[\alpha_j] = \epsilon^{-1}$ or $\mathbb{E}_{\mathcal{Q}}[\beta_k] = \epsilon^{-1}$ respectively.
In the limit as $\epsilon \to 0$, using Equation (22) and block matrix inversion formula we get

$$\begin{aligned}\Sigma_{W_{jk}} &\leftarrow \lim_{\epsilon\to 0}\left(\left(\mathbb{E}_{\mathcal{Q}_{/W}}[\mathrm{diag}(\beta)\otimes\mathrm{diag}(\alpha) + \sigma^{-2}\sum_i\Big((\Psi_i)\otimes(Z_i^T Z_i)\Big)]\right)^{-1}\right)_{[j+kM,j+kM]} \\ &= \lim_{\epsilon\to 0}\left(\left(\mathbb{E}_{\mathcal{Q}_{/W}}[\alpha_j\beta_k]\right)^{-1} + \mathcal{O}(\epsilon^2)\right) = \left(\mathbb{E}_{\mathcal{Q}_{/W}}[\alpha_j\beta_k]\right)^{-1} = \frac{b_{\alpha_j}b_{\beta_k}}{a_{\alpha_j}a_{\beta_k}}\end{aligned} \tag{83}$$

In the limit as $\epsilon \to 0$ and using Equation (23)

$$\begin{aligned}\mu_{W_{jk}} &\leftarrow \lim_{\epsilon\to 0}\left(-\frac{a_\sigma}{b_\sigma}\sum_i \mathrm{vec}\left(\left(\Phi_i(\mu_{\bar{Z}}\Phi_i - Y_i)^T\mu_{Z_i}\right)^T\right)^T\Sigma_{\mathrm{vec}(W)}\right)_{[1,j+kM]} \\ &= \left(-\frac{a_\sigma}{b_\sigma}\sum_i \mathrm{vec}\left(\left(\Phi_i(\mu_{\bar{Z}}\Phi_i - Y_i)^T\mu_{Z_i}\right)^T\right)^T\right)\left(\Sigma_{\mathrm{vec}(W)}\right)_{\cdot(j+kM)} \in \mathcal{O}(\epsilon)\end{aligned} \tag{84}$$

 doi:10.20944/preprints202503.0658.v1

Equation (84) uses the fact that elements in $\left(\Sigma_{\text{vec}(W)}\right)_{\cdot(j+kM)}$ are all $\mathcal{O}(\epsilon)$ based on block matrix inversion formula. Thus,

$$\begin{aligned}\lim_{\epsilon\to 0}\mathbb{E}_{\mathcal{Q}}[W_{jk}^2] &= \lim_{\epsilon\to 0}\left(\Sigma_{W_{jk}} + (\mu_{W_{jk}})^2\right) = \lim_{\epsilon\to 0}\left(\Sigma_{W_{jk}} + \mathcal{O}(\epsilon^2)\right)\\ &= \lim_{\epsilon\to 0}\left(\frac{b_{\alpha_j}b_{\beta_k}}{a_{\alpha_j}a_{\beta_k}} + \mathcal{O}(\epsilon^2)\right) = \frac{b_{\alpha_j}b_{\beta_k}}{a_{\alpha_j}a_{\beta_k}}\end{aligned} \tag{85}$$

□

Lemma 6. *$\forall k > K^{(a)}$, after updating $\mathcal{Q}_{\bar{Z}_B}$ by Equations (20) and (21), $\mathbb{E}_{\mathcal{Q}}[\bar{Z}_k^2] = \frac{b_\eta}{a_\eta}\epsilon$.*

Proof. If $k > K^{(a)}$, $\mathbb{E}_{\mathcal{Q}}[\beta_k] = \epsilon^{-1}$.
Then using Equation (20) and block matrix inversion formula we get

$$\begin{aligned}\Sigma_{\bar{Z}kk} &\leftarrow \lim_{\epsilon\to 0}\left(\left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sum_{i=1}^{P}\left(\sigma^{-2}\Psi_i\right) + \eta\,\text{diag}(\beta)\right]\right)^{-1}\right)_{kk}\\ &= \lim_{\epsilon\to 0}\left(\left(\sum_{i=1}^{P}\left(\frac{a_\sigma}{b_\sigma}\Psi_i\right) + \frac{a_\eta}{b_\eta}\,\text{diag}(\frac{a}{b})\right)^{-1}\right)_{kk}\\ &= \lim_{\epsilon\to 0}\left(\frac{b_\eta b_{\beta_k}}{a_\eta a_{\beta_k}} + \mathcal{O}(\epsilon^2)\right) = \frac{b_\eta b_{\beta_k}}{a_\eta a_{\beta_k}}\end{aligned} \tag{86}$$

Using Equation (21)

$$\begin{aligned}\mu_{\bar{Z}k} &\leftarrow \lim_{\epsilon\to 0}\left(\left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\right]\sum_{i=1}^{P}(Y - \mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[Z_i W]\Phi_i)\Phi_i^T\right)\Sigma_{\bar{Z}}\right)_{1k}\\ &= \lim_{\epsilon\to 0}\left(\left(\frac{a_\sigma}{b_\sigma}\sum_{i=1}^{P}(Y - \mu_{Z_i}\mu_W\Phi_i)\Phi_i^T\right)\Sigma_{\bar{Z}}\right)_{1k}\\ &= \lim_{\epsilon\to 0}\left(\frac{a_\sigma}{b_\sigma}\sum_{i=1}^{P}(Y - \mu_{Z_i}\mu_W\Phi_i)\Phi_i^T\right)\Sigma_{\bar{Z}\cdot k} \in \mathcal{O}(\epsilon)\end{aligned} \tag{87}$$

Equation (87) uses the fact that elements in $\Sigma_{\bar{Z}\cdot k}$ are all $\in \mathcal{O}(\epsilon)$.

$$\begin{aligned}\mathbb{E}_{\mathcal{Q}}[\bar{Z}_k^2] &= \lim_{\epsilon\to 0}\left(\Sigma_{\bar{Z}kk} + \mu_{\bar{Z}k}^2\right) = \lim_{\epsilon\to 0}\left(\Sigma_{\bar{Z}kk} + \mathcal{O}(\epsilon^2)\right)\\ &= \lim_{\epsilon\to 0}\left(\frac{b_\eta b_{\beta_k}}{a_\eta a_{\beta_k}} + \mathcal{O}(\epsilon^2)\right) = \lim_{\epsilon\to 0}\left(\frac{b_\eta}{a_\eta}\epsilon + \mathcal{O}(\epsilon^2)\right) = \frac{b_\eta}{a_\eta}\epsilon\end{aligned} \tag{88}$$

□

Theorem 1. *$\forall j \le J^{(a)}$, updates of $\mathcal{Q}_{\alpha_j}$ and $\mathcal{Q}_{W_B}$ will converge at $\mathbb{E}_{\mathcal{Q}_{\alpha_j}}[\alpha_j] = \mathbb{E}_{\mathcal{Q}_{\alpha_j}^{(a)}}[\alpha_j]$ given $\mathbb{E}_{\mathcal{Q}_{\beta_k}}[\beta_k] = \mathbb{E}_{\mathcal{Q}_{\beta_k}^{(a)}}[\beta_k], \forall k \le K^{(a)}$, $a_0 = b_0 = 0$ and conditions in Equations (80)-(79) are satisfied in the limit as $\epsilon \to 0$.*

 doi:10.20944/preprints202503.0658.v1

Proof. Assume $\mathcal{Q}_{\alpha_j}^{(\mathrm{a})}$ has just been updated using Equations (14), (15), i.e., $\forall j \leq J^{(\mathrm{a})}$

$$a_{\alpha_j}^{(\mathrm{a})} = a_0 + \frac{K^{(\mathrm{a})}}{2} \tag{89}$$

$$\begin{aligned} b_{\alpha_j}^{(\mathrm{a})} &= b_0 + \frac{1}{2}\sum_{k=1}^{K^{(\mathrm{a})}} \mathbb{E}_{\mathcal{Q}_{/\alpha_j}^{(\mathrm{a})}}[W_{jk}^2 \beta_k] \\ &= b_0 + \frac{1}{2}\sum_{k=1}^{K^{(\mathrm{a})}} \left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right) \frac{a_{\beta_k}^{(\mathrm{a})}}{b_{\beta_k}^{(\mathrm{a})}} \right) \end{aligned} \tag{90}$$

The updates for $\mathcal{Q}_\alpha$ derived from $\mathcal{L}$ are

$$\begin{aligned} b_{\alpha_j} &\leftarrow b_0 + \frac{1}{2}\sum_{k=1}^{K} \left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \\ &= b_0 + \frac{1}{2}\sum_{k=1}^{K^{(\mathrm{a})}} \left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \\ &+ \frac{1}{2}\sum_{k=K^{(\mathrm{a})}+1}^{K} \left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \\ &= b_{\alpha_j}^{(\mathrm{a})} + \frac{1}{2}\sum_{k=K^{(\mathrm{a})}+1}^{K} \left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \end{aligned} \tag{91}$$

It involves $W_{jk}, k > K^{(\mathrm{a})}$ and therefore they need to be kepted updated. Apply Theorem 5 for Equation (91) and we can get

$$\begin{aligned} b_{\alpha_j} &\leftarrow b_{\alpha_j}^{(\mathrm{a})} + \frac{1}{2}\sum_{k=K^{(\mathrm{a})}+1}^{K} \left(\left(\frac{b_{\alpha_j} b_{\beta_k}}{a_{\alpha_j} a_{\beta_k}}\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \\ &= b_{\alpha_j}^{(\mathrm{a})} + \frac{1}{2}(K - K^{(\mathrm{a})}) \frac{b_{\alpha_j}}{a_{\alpha_j}} \end{aligned} \tag{92}$$

Applying Equation (92) in an iterative manner, we will get a sequence of updates for a_{α_j}. Solving

$$b_{\alpha_j} = b_{\alpha_j}^{(\mathrm{a})} + \frac{1}{2}(K - K^{(\mathrm{a})}) \frac{b_{\alpha_j}}{\frac{K}{2}} \tag{93}$$

$$\Rightarrow b_{\alpha_j} = (1 - \frac{1}{2}(K - K^{(\mathrm{a})}) \frac{2}{K})^{-1} b_{\alpha_j}^{(\mathrm{a})} = \frac{K}{K^{(\mathrm{a})}} b_{\alpha_j}^{(\mathrm{a})} \tag{94}$$

Thus, we can get that the sequence will converge at

$$b_{\alpha_j} \leftarrow \frac{K}{K^{(\mathrm{a})}} b_{\alpha_j}^{(\mathrm{a})} \tag{95}$$

As a result, $\mathbb{E}_{\mathcal{Q}_{\alpha_j}}[\alpha_j] = \frac{a_{\alpha_j}}{b_{\alpha_j}} = \frac{a_{\alpha_j}^{(\mathrm{a})}}{b_{\alpha_j}^{(\mathrm{a})}} = \mathbb{E}_{\mathcal{Q}_{\alpha_j}^{(\mathrm{a})}}[\alpha_j]$. □

Theorem 2. *$\forall k \leq K^{(a)}$, updates of $\mathcal{Q}_{\beta_k}$ and $\mathcal{Q}_{W_C}$ will converge at $\mathbb{E}_{\mathcal{Q}_{\beta_k}}[\beta_k] = \mathbb{E}_{\mathcal{Q}_{\beta_k}^{(a)}}[\beta_k]$ given $\mathbb{E}_{\mathcal{Q}_{\alpha_j}}[\alpha_j] = \mathbb{E}_{\mathcal{Q}_{\alpha_j}^{(a)}}[\alpha_j], \forall j \leq J^{(a)}$, $a_0 = b_0 = 0$ and conditions in Equations (80)-(79) are satisfied in the limit as $\epsilon \to 0$.*

 doi:10.20944/preprints202503.0658.v1

Proof. Assume $\mathcal{Q}^{(\mathrm{a})}_{\beta_k}$ has just been updated using Equations (18) and (19), i.e.,

$$a^{(\mathrm{a})}_{\beta_k} = a_0 + \frac{K^{(\mathrm{a})}+1}{2} \tag{96}$$

$$\begin{aligned} b^{(\mathrm{a})}_{\beta_k} &\leftarrow b_0 + \frac{1}{2}\mathbb{E}_{\mathcal{Q}^{(\mathrm{a})}_{/\beta_k}}[\bar{Z}_k^2 + \sum_{j=1}^{J^{(\mathrm{a})}} W_{jk}^2 \alpha_j] \\ &= b_0 + \frac{1}{2}\left(\left(\Sigma_{\bar{Z}kk} + \mu_{\bar{Z}k}^2\right) + \sum_{j=1}^{J^{(\mathrm{a})}}\left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right)\frac{a^{(\mathrm{a})}_{\alpha_j}}{b^{(\mathrm{a})}_{\alpha_j}}\right)\right) \end{aligned} \tag{97}$$

The updates for $\mathcal{Q}_{\beta_k}$ derived from $\mathcal{L}$ is

$$\begin{aligned} b_{\beta_k} &\leftarrow b_0 + \frac{1}{2}\left(\left(\Sigma_{\bar{Z}kk} + \mu_{\bar{Z}k}^2\right) + \sum_{j=1}^{J}\left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right)\frac{a_{\alpha_j}}{b_{\alpha_j}}\right)\right) \\ &= b_0 + \frac{1}{2}\left(\left(\Sigma_{\bar{Z}kk} + \mu_{\bar{Z}k}^2\right) + \sum_{j=1}^{J^{(\mathrm{a})}}\left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right)\frac{a_{\alpha_j}}{b_{\alpha_j}}\right)\right) \\ &+ \frac{1}{2}\left(\sum_{j=J^{(\mathrm{a})}+1}^{J}\left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right)\frac{a_{\alpha_j}}{b_{\alpha_j}}\right)\right) \\ &= b^{(\mathrm{a})}_{\beta_k} + \frac{1}{2}\left(\sum_{j=J^{(\mathrm{a})}+1}^{J}\left(\left(\Sigma_{W_{jk}} + \mu_{W_{jk}}^2\right)\frac{a_{\alpha_j}}{b_{\alpha_j}}\right)\right) \end{aligned} \tag{98}$$

It involves $W_{jk}, j > J^{(\mathrm{a})}$ and therefore they need to be kepted updated. Apply Theorem 5 for Equation (98) and we can get

$$b_{\beta_k} \leftarrow b^{(\mathrm{a})}_{\beta_k} + \frac{1}{2}\sum_{j=J^{(\mathrm{a})}+1}^{J}\left(\left(\frac{b_{\beta_k} b_{\alpha_j}}{a_{\beta_k} a_{\alpha_j}}\right)\frac{a_{\alpha_j}}{b_{\alpha_j}}\right) \tag{99}$$

$$\Rightarrow = b^{(\mathrm{a})}_{\beta_k} + \frac{1}{2}(K - K^{(\mathrm{a})})\frac{b_{\beta_k}}{a_{\beta_k}} \tag{100}$$

Applying Equation (100) in an iterative manner, we will get a sequence of b_{β_k}. Solving

$$b_{\beta_k} = b^{(\mathrm{a})}_{\beta_k} + \frac{1}{2}(K - K^{(\mathrm{a})})\frac{b_{\beta_k}}{\frac{K+1}{2}} \tag{101}$$

$$b_{\beta_k} = (1 - \frac{1}{2}(K - K^{(\mathrm{a})})\frac{2}{K+1})^{-1} b^{(\mathrm{a})}_{\beta_k} = \frac{K+1}{K^{(\mathrm{a})}+1} b^{(\mathrm{a})}_{\beta_k} \tag{102}$$

Thus, we can get that the sequence will converge at

$$b_{\beta_k} \leftarrow \frac{K+1}{K^{(\mathrm{a})}+1} b^{(\mathrm{a})}_{\beta_k} \tag{103}$$

As a result, $\mathbb{E}_{\mathcal{Q}}[\beta_k] = \frac{a_{\beta_k}}{b_{\beta_k}} = \frac{a^{(\mathrm{a})}_{\beta_k}}{b^{(\mathrm{a})}_{\beta_k}} = \mathbb{E}_{\mathcal{Q}^{(\mathrm{a})}}[\beta_k]$. □

Theorem 3. *Updates of $\mathcal{Q}_\eta$ and $\mathcal{Q}_{\bar{Z}_B}$ will converge at $\mathbb{E}_{\mathcal{Q}_\eta}[\eta] = \mathbb{E}_{\mathcal{Q}^{(a)}_\eta}[\eta]$ given $\mathbb{E}_{\mathcal{Q}_{\beta_k}}[\beta_k] = \mathbb{E}_{\mathcal{Q}^{(a)}_{\beta_k}}[\beta_k], \forall k \leq K^{(a)}$, $a_0 = b_0 = 0$ and conditions in Equations (80)-(79) are satisfied in the limit as $\epsilon \to 0$.*

 doi:10.20944/preprints202503.0658.v1

Proof. Assume $\mathcal{Q}_\eta^{(\mathrm{a})}$ has just been updated using Equations (16) and (17), i.e.,

$$a_\eta^{(\mathrm{a})} \leftarrow a_0 + \frac{K^{(\mathrm{a})}}{2} \tag{104}$$

$$\begin{aligned} b_\eta{}^{(\mathrm{a})} &\leftarrow b_0 + \frac{1}{2}\sum_{k=1}^{K^{(\mathrm{a})}} \mathbb{E}_{\mathcal{Q}_{/\eta}}[\bar{Z}_k^2 \beta_k] \\ &= b_0 + \frac{1}{2}\sum_{k=1}^{K^{(\mathrm{a})}} \left(\left(\Sigma_{\bar{Z}k} + \mu_{\bar{Z}k}^2\right) \frac{a_{\beta_k}^{(\mathrm{a})}}{b_{\beta_k}^{(\mathrm{a})}} \right) \end{aligned} \tag{105}$$

The updates for $\mathcal{Q}_\eta$ derived from $\mathcal{L}$ is

$$\begin{aligned} b_\eta &\leftarrow b_0 + \frac{1}{2}\sum_{k=1}^{K} \left(\left(\Sigma_{\bar{Z}k} + \mu_{\bar{Z}k}^2\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \\ &= b_0 + \frac{1}{2}\sum_{k=1}^{K^{(\mathrm{a})}} \left(\left(\Sigma_{\bar{Z}k} + \mu_{\bar{Z}k}^2\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) + \frac{1}{2}\sum_{k=K^{(\mathrm{a})}+1}^{K} \left(\left(\Sigma_{\bar{Z}k} + \mu_{\bar{Z}k}^2\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \\ &= b_\eta{}^{(\mathrm{a})} + \frac{1}{2}\sum_{k=K^{(\mathrm{a})}+1}^{K} \left(\left(\Sigma_{\bar{Z}k} + \mu_{\bar{Z}k}^2\right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \end{aligned} \tag{106}$$

It involves $\bar{Z}_k, k > K^{(\mathrm{a})}$ and therefore they need to be kept updated. Apply Lemma 6 for Equation (106) and we can get

$$b_\eta \leftarrow b_\eta{}^{(\mathrm{a})} + \frac{1}{2}\sum_{k=K^{(\mathrm{a})}+1}^{K} \left(\left(\frac{b_\eta b_{\beta_k}}{a_\eta a_{\beta_k}} \right) \frac{a_{\beta_k}}{b_{\beta_k}} \right) \tag{107}$$

$$= b_\eta{}^{(\mathrm{a})} + \frac{1}{2}(K - K^{(\mathrm{a})}) \frac{b_\eta}{a_\eta} \tag{108}$$

Applying Equation (108) in an iterative manner, we will get a sequence of updates for b_η. Solving

$$b_\eta = b_\eta{}^{(\mathrm{a})} + \frac{1}{2}(K - K^{(\mathrm{a})}) \frac{b_\eta}{\frac{K}{2}} \tag{109}$$

$$b_\eta = (1 - \frac{1}{2}(K - K^{(\mathrm{a})}) \frac{2}{K})^{-1} b_\eta{}^{(\mathrm{a})} = \frac{K}{K^{(\mathrm{a})}} b_\eta{}^{(\mathrm{a})} \tag{110}$$

Thus, we can get that the sequence will converge at

$$b_\eta \leftarrow \frac{K}{K^{(\mathrm{a})}} b_\eta{}^{(\mathrm{a})} \tag{111}$$

As a result, $\mathbb{E}_{\mathcal{Q}}[\eta] = \frac{b_\eta}{a_\eta} = \frac{b_\eta{}^{(\mathrm{a})}}{a_\eta^{(\mathrm{a})}} = \mathbb{E}_{\mathcal{Q}^{(\mathrm{a})}}[\eta]$. □

In practice, due to limitations in numerical representation, we restrict values so that the active precision parameter estimates would not really go to infinity:

$$\mathbb{E}_{\mathcal{Q}_{\alpha_j}}[\alpha_j] \leq \tau_{\max}, \forall j \leq J^{(\mathrm{a})} \tag{112}$$

$$\mathbb{E}_{\mathcal{Q}_{\beta_k}}[\beta_k] \leq \tau_{\max}, \forall k \leq K^{(\mathrm{a})} \tag{113}$$

 doi:10.20944/preprints202503.0658.v1

9.3. Weights and Noise

Here is how to update $\mathcal{Q}_{Z_A}, \mathcal{Q}_{\bar{Z}_A}, \mathcal{Q}_{W_A}, \mathcal{Q}_\sigma$ in a scalable manner, using computation in the $K^{(a)}$ dimension subspace only.

Theorem 4. *$\mathcal{L}$ and $\mathcal{L}^{(a)}$ share the same update rule for Z_{iA}, i.e.,*

$$\begin{aligned} H_{iAjk} &\leftarrow \mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_{Aj}\Phi_{iA}\Phi_{iA}^T W_{Ak}^T] = \mathrm{Tr}(\mathbb{E}_{\mathcal{Q}_{/Z_i}}[W_{Ak}^T W_{Aj}]\Phi_{iA}\Phi_{iA}^T) \\ &= Tr\Big(\Big(\Sigma_{[W_{Ak},W_{Aj}]} + \mu_{[W_{Aj}]}^T \mu_{[W_{Ak}]}\Big)\Phi_{iA}\Phi_{iA}^T\Big), \forall j = 1:J^{(a)}, k = 1:K^{(a)} \end{aligned} \tag{114}$$

$$\Sigma_{Z_i A} \leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/Z_i}}[\sigma^{-2} W_A \Phi_{iA}\Phi_{iA}^T W_A^T + I]\right)^{-1} = [\frac{a_\sigma}{b_\sigma} H_{iA} + I]^{-1} \tag{115}$$

$$\mu_{iA} \leftarrow \mathbb{E}_{\mathcal{Q}_{/Z_i}}[\sigma^{-2}(Y_i - \bar{Z}\Phi_{iA})\Phi_{iA}^T W_A^T]\Sigma_{Z_i A} = \frac{a_\sigma}{b_\sigma}(Y_i - \mu_{\bar{Z}A}\Phi_{iA})\Phi_{iA}^T(\mu_{W_A})^T \Sigma_{Z_i A} \tag{116}$$

Proof. Apply Lemma 3 to Equation (24) we get

$$\begin{aligned} H_{iAjk} &\leftarrow Tr\Big(\Big(\Sigma_{[W_{Ak},W_{Aj}]} + \mu_{[W_{Aj}]}^T \mu_{[W_{Ak}]}\Big)\Phi_{iA}\Phi_{iA}^T\Big) + \mathcal{O}(\epsilon) \\ &\to Tr\Big(\Big(\Sigma_{[W_{Ak},W_{Aj}]} + \mu_{[W_{Aj}]}^T \mu_{[W_{Ak}]}\Big)\Phi_{iA}\Phi_{iA}^T\Big), \forall j = 1:J^{(a)}, k = 1:K^{(a)} \end{aligned} \tag{117}$$

Apply block matrix inversion formula to Equation (25) we get

$$\begin{aligned} \Sigma_{Z_i A} &\leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/Z_i}}[\sigma^{-2} W_A \Phi_{iA}\Phi_{iA}^T W_A^T + I]\right)^{-1} + \mathcal{O}(\epsilon^2) \\ &\to \left(\mathbb{E}_{\mathcal{Q}_{/Z_i}}[\sigma^{-2} W_A \Phi_{iA}\Phi_{iA}^T W_A^T + I]\right)^{-1} = [\frac{a_\sigma}{b_\sigma} H_{iA} + I]^{-1} \end{aligned} \tag{118}$$

Apply block matrix multiplication and Theorem 5 to Equation (26) conditioned on Equation (81) we get

$$\begin{aligned} \mu_{iA} &\leftarrow \mathbb{E}_{\mathcal{Q}_{/Z_i}}[\sigma^{-2}(Y_i - \bar{Z}\Phi_{iA})\Phi_{iA}^T W_A^T]\Sigma_{Z_i A} + \mathcal{O}(\epsilon) \\ &\to \frac{a_\sigma}{b_\sigma}(Y_i - \mu_{\bar{Z}A}\Phi_{iA})\Phi_{iA}^T(\mu_{W_A})^T \Sigma_{Z_i A} \end{aligned} \tag{119}$$

□

Theorem 5. *$\mathcal{L}$ and $\mathcal{L}^{(a)}$ share the same update rule for $\bar{Z}_A$, i.e.,*

$$\begin{aligned} \Sigma_{\bar{Z}A} &\leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sum_{i=1}^P \left(\sigma^{-2}\Phi_{iA}\Phi_{iA}^T\right) + \eta\,\mathrm{diag}(\beta_A)\right]\right)^{-1} \\ &= \left(\sum_{i=1}^P \left(\frac{a_\sigma}{b_\sigma}\Phi_{iA}\Phi_{iA}^T\right) + \frac{a_\eta}{b_\eta}\,\mathrm{diag}(\frac{a_A}{b_A})\right)^{-1} \end{aligned} \tag{120}$$

$$\begin{aligned} \mu_{\bar{Z}A} &\leftarrow \left(\mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\right]\sum_{i=1}^P (Y - \mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}[Z_{iA}W_A]\Phi_{iA})\Phi_{iA}\right)\Sigma_{\bar{Z}A} \\ &= \left(\frac{a_\sigma}{b_\sigma}\sum_{i=1}^P (Y - \mu_{iA}\mu_{W_A}\Phi_{iA})\Phi_{iA}\right)\Sigma_{\bar{Z}A} \end{aligned} \tag{121}$$

 doi:10.20944/preprints202503.0658.v1

Proof. Apply block matrix inversion formula to Equation (20) conditioned on $\mathbb{E}_{\mathcal{Q}/\bar{\mathcal{Z}}}[\beta_k] = \epsilon^{-1}, \forall k > K^{(a)}$ and we get

$$\begin{aligned}\Sigma_{\bar{Z}A} &\leftarrow \left(\mathbb{E}_{\mathcal{Q}/\bar{Z}}\left[\sum_{i=1}^{P}\left(\sigma^{-2}\Phi_{iA}\Phi_{iA}^T\right) + \eta\,\mathrm{diag}(\beta_A)\right]\right)^{-1} + \mathcal{O}(\epsilon)\\ &\to \left(\mathbb{E}_{\mathcal{Q}/\bar{Z}}\left[\sum_{i=1}^{P}\left(\sigma^{-2}\Phi_{iA}\Phi_{iA}^T\right) + \eta\,\mathrm{diag}(\beta_A)\right]\right)^{-1}\end{aligned} \tag{122}$$

Apply block matrix multiplication and Theorem 5 to Equation (21) conditioned on Equation (80) we get

$$\begin{aligned}\mu_{\bar{Z}A} &\leftarrow \left(\mathbb{E}_{\mathcal{Q}/\bar{Z}}\left[\sigma^{-2}\right]\sum_{i=1}^{P}(Y - \mathbb{E}_{\mathcal{Q}/\bar{Z}}[Z_{iA}W_A]\Phi_{iA})\Phi_{iA}\right)\Sigma_{\bar{Z}A} + \mathcal{O}(\epsilon)\\ &\to \left(\mathbb{E}_{\mathcal{Q}/\bar{Z}}\left[\sigma^{-2}\right]\sum_{i=1}^{P}(Y - \mathbb{E}_{\mathcal{Q}/\bar{Z}}[Z_{iA}W_A]\Phi_{iA})\Phi_{iA}\right)\Sigma_{\bar{Z}A}\end{aligned} \tag{123}$$

□

Theorem 6. *$\mathcal{L}$ and $\mathcal{L}^{(a)}$ share the same update rule for W_A, i.e.,*

$$\begin{aligned}\Sigma_{\mathrm{vec}(W)} &\leftarrow \mathbb{E}_{\mathcal{Q}/W}\left[\sigma^{-2}\sum_{i=1}^{P}\left((\Phi_{iA}^T\Phi_{iA})\otimes(Z_{iA}^TZ_{iA})\right) + \mathrm{diag}(\beta_A)\otimes\mathrm{diag}(\alpha_A)\right]^{-1}\\ &= \left(\frac{a_\sigma}{b_\sigma}\sum_{i=1}^{P}\left((\Phi_{iA}^T\Phi_{iA})\otimes(\mu_{iA}^T\mu_{iA} + \Sigma_{Z_iA})\right) + \mathrm{diag}\left(\frac{a_A}{b_A}\right)\otimes\mathrm{diag}\left(\frac{c_A}{d_A}\right)\right)^{-1}\end{aligned} \tag{124}$$

$$\begin{aligned}\mu_{\mathrm{vec}(W)} &\leftarrow \mathbb{E}_{\mathcal{Q}/W}\left[-\sigma^{-2}\sum_{i=1}^{P}\mathrm{vec}\left(\left(\Phi_{iA}(\Phi_{iA}^T\bar{Z}_A^T - Y_i^T)Z_{iA}\right)^T\right)^T\right]\Sigma_{\mathrm{vec}(W)A}\\ &= -\frac{a_\sigma}{b_\sigma}\sum_{i=1}^{P}\mathrm{vec}\left(\left(\Phi_{iA}(\Phi_{iA}^T\mu_{\bar{Z}A}^T - Y_i^T)\mu_{iA}\right)^T\right)^T\Sigma_{\mathrm{vec}(W)A}\end{aligned} \tag{125}$$

Proof. Apply block matrix inversion formula to Equation (22) conditioned on $\mathbb{E}_{\mathcal{Q}/\bar{\mathcal{Z}}}[\beta_k] = \epsilon^{-1}, \forall k > K^{(a)}$ and $\mathbb{E}_{\mathcal{Q}/\bar{\mathcal{Z}}}[\alpha_j] = \epsilon^{-1}, \forall j > J^{(a)}$, we get

$$\begin{aligned}\Sigma_{\mathrm{vec}(W)} &\leftarrow \mathbb{E}_{\mathcal{Q}/W}\left[\sigma^{-2}\sum_{i=1}^{P}\left((\Phi_{iA}^T\Phi_{iA})\otimes(Z_{iA}^TZ_{iA})\right) + \mathrm{diag}(\beta_A)\otimes\mathrm{diag}(\alpha_A)\right]^{-1} + \mathcal{O}(\epsilon)\\ &\to \mathbb{E}_{\mathcal{Q}/W}\left[\sigma^{-2}\sum_{i=1}^{P}\left((\Phi_{iA}^T\Phi_{iA})\otimes(Z_{iA}^TZ_{iA})\right) + \mathrm{diag}(\beta_A)\otimes\mathrm{diag}(\alpha_A)\right]^{-1}\end{aligned} \tag{126}$$

Apply block matrix multiplication and Theorem 5 to Equation (23) conditioned on Equation (80) and Equation (81), we get

$$\begin{aligned}\mu_{\mathrm{vec}(W)} &\leftarrow \mathbb{E}_{\mathcal{Q}/W}\left[-\sigma^{-2}\sum_{i=1}^{P}\mathrm{vec}\left(\left(\Phi_{iA}(\Phi_{iA}^T\bar{Z}_A^T - Y_i^T)Z_{iA}\right)^T\right)^T\right]\Sigma_{\mathrm{vec}(W)A} + \mathcal{O}(\epsilon)\\ &\to \mathbb{E}_{\mathcal{Q}/W}\left[-\sigma^{-2}\sum_{i=1}^{P}\mathrm{vec}\left(\left(\Phi_{iA}(\Phi_{iA}^T\bar{Z}_A^T - Y_i^T)Z_{iA}\right)^T\right)^T\right]\Sigma_{\mathrm{vec}(W)A}\end{aligned} \tag{127}$$

□

 doi:10.20944/preprints202503.0658.v1

Theorem 7. *$\mathcal{L}$ and $\mathcal{L}^{(a)}$ share the same update rule for σ, i.e.,*

$$a_\sigma \leftarrow a_0 + \frac{1}{2}\sum_i N_i \tag{128}$$

$$\begin{aligned} b_\sigma &\leftarrow b_0 + \frac{1}{2}\mathbb{E}_{\mathcal{Q}_{/\sigma}}\left[\sum_i ||Y_i - (Z_{iA}W_A + \bar{Z}_A)\Phi_{iA}||_2^2\right] \\ &= b_0 + \frac{1}{2}\sum_i (Y_iY_i^T - 2Y_i(\mu_{iA}\mu_{W_A}\Phi_{iA})^T - 2Y_i(\mu_{\bar{Z}A}\Phi_{iA})^T + 2\mu_{iA}\mu_{W_A}\Phi_{iA}\Phi_{iA}^T(\mu_{\bar{Z}A})^T \\ &+ \mathrm{Tr}\Big(\Big(\Sigma_{\bar{Z}A} + (\mu_{\bar{Z}A})^T\mu_{\bar{Z}A}\Big)\Phi_{iA}\Phi_{iA}^T\Big)) \\ &+ \frac{1}{2}\mathrm{vec}(G_A^T)^T\sum_i \mathrm{vec}\Big(\mathrm{vec}(\Phi_{iA}\Phi_{iA}^T)\,\mathrm{vec}(\Sigma_{Z_iA} + \mu_{iA}^T\mu_{iA})^T\Big), \end{aligned} \tag{129}$$

where

$$\begin{aligned} G_{A(j+kM)} &\leftarrow \mathbb{E}_{\mathcal{Q}_{/\sigma}}\left[\mathrm{vec}(W_{Ak}W_{Aj}^T)^T\right] \\ &= \mathrm{vec}(\Sigma_{[W_{Ak},W_{Aj}]} + \mu_{\mathrm{vec}(W)_{[W_{Aj}]}}^T\mu_{\mathrm{vec}(W)_{[W_{Ak}]}})^T, \forall j = 1:K^{(a)}, k = 1:K^{(a)} \end{aligned} \tag{130}$$

Proof. Apply block matrix multiplication and Theorem 5 to Equation (28) conditioned on Equations (80) and (81), we get

$$\begin{aligned} b_\sigma &\leftarrow b_0 + \frac{1}{2}\mathbb{E}_{\mathcal{Q}_{/\sigma}}\left[\sum_i ||Y_i - (Z_{iA}W_A + \bar{Z}_A)\Phi_{iA}||_2^2\right] + \mathcal{O}(\epsilon) \\ &\to b_0 + \frac{1}{2}\mathbb{E}_{\mathcal{Q}_{/\sigma}}\left[\sum_i ||Y_i - (Z_{iA}W_A + \bar{Z}_A)\Phi_{iA}||_2^2\right] \end{aligned} \tag{131}$$

□

We show $\mathcal{Q}_{Z_{iA}}, \mathcal{Q}_{W_A}, \mathcal{Q}_{\bar{Z}_{iA}}, \mathcal{Q}_\sigma$ share the same update formulas as those derived from the low-dimensional lower bound: $\mathcal{Q}^{(a)}{}_{Z_{iA}}, \mathcal{Q}^{(a)}{}_{W_A}, \mathcal{Q}^{(a)}{}_{\bar{Z}_{iA}}, \mathcal{Q}^{(a)}{}_\sigma$. Thus, in practice, if suffices to update $\mathcal{Q}^{(a)}$; we can then increase $K^{(a)}$ by including new basis functions. This process proves to implicitly maximizes $\mathcal{L}$ with $\mathcal{Q}$.

9.4. Low-dimensional Lower Bound

We now have updating formulas for the parameters in the active subspace. $\mathcal{Q}_{Z_{iA}}$ is updated by Equations (114), (115) and (116). $\mathcal{Q}_{W_A}$ is updated by Equations (124) and (125). $\mathcal{Q}_{\bar{Z}_A}$ is updated by Equations (120) and (121). $\mathcal{Q}_{\alpha_A}, \mathcal{Q}_{\beta_A}, \mathcal{Q}_\eta$ are updated by Theorem 1, 2 and 3, with the companion of implicit updates of $\mathcal{Q}_{W_B}, \mathcal{Q}_{W_C}$. $\mathcal{Q}_\sigma$ is updated by Equations (130), (128) and (129). All the updating rules are identical to those derived from the low-dimensional lower bound $\mathcal{L}^{(a)}$ with $K^{(a)}$ basis functions. Therefore, in practice all we need is to optimize $\mathcal{L}^{(a)}$, with time complexity of $\mathcal{O}\Big({K^{(a)}}^2 \max\Big({K^{(a)}}^4, P\max_i(N_i)\Big)\Big)$, as described in Theorem 8, and then check if a new basis function should be included in the model.

For numerical stability, we scale ϕ, b such that $\min_k(\mathbb{E}_{Q_\beta}[\beta_k]) = \min_k(\frac{c_k}{d_k}) = 1$ at the beginning of Algorithm 2.

Theorem 8. *The lower bound $\mathcal{L}$ can be optimized using Algorithm 2 with time complexity of $\mathcal{O}\Big({K^{(a)}}^2 \max\Big({K^{(a)}}^4, P\max_i(N_i)\Big)\Big)$.*

Proof. It is a consequence of Theorems 1, 2, 3, 4, 5, 6, 7. □

 doi:10.20944/preprints202503.0658.v1

Algorithm 2 Variational inference

Require: $\mu_{Z_i}, \Sigma_{Z_i}, \mu_{\mathrm{vec}(W)}, \Sigma_{\mathrm{vec}(W)}, \mu_{\bar{Z}}, \Sigma_{\bar{Z}}, a_\sigma, b_\sigma, a_{\alpha_j}, b_{\alpha_j}, a_{\beta_k}, b_{\beta_k}, \forall i, j, k$ ▷ Multi sample RVM
  while True **do**
    $\mathcal{L}^{(a)} \leftarrow \mathrm{lowerbound}(\mathcal{Q}^{(a)})$
    Update $\mathcal{Q}^{(a)}$ with respect to all parameters using mean field approximation
    if $\mathrm{lowerbound}(\mathcal{Q}^{(a)}) - \mathcal{L}^{(a)} < \tau_{\mathrm{con}}$ **then** ▷ Insignificant increase
      Search for new basis functions using Algorithm 1
      if not found **then** ▷ Converged
        break
      end if
    end if
    Remove dimensions associated with the precision of the maximum values
  end while
  Get rid of dimensions associated with $\alpha_j \geq \min_j(\alpha_j)\tau_{\mathrm{eff}}$ and $\beta_k \geq \min_k(\beta_k)\tau_{\mathrm{eff}}$

10. Scalable Update for BSFDA$^{\mathbf{Fast}}$

For brevity, we denote the covariance of ζ_i as S, i.e., $\zeta_i \sim \mathcal{N}(0, S)$. S is diagonal and $S_{kk} = \varsigma_k^2 \beta_k^{-1}$. The variational update formulas are as follows:

$$\Sigma_{\theta_i} \leftarrow \mathbb{E}_{\mathcal{Q}_{/\theta_i}}\left[\Phi_i \Phi_i^T \sigma^{-2} + S^{-1}\right]^{-1} \tag{132}$$

$$\mu_{\theta_i} \leftarrow \mathbb{E}_{\mathcal{Q}_{/\theta_i}}\left[\left((Y_i - \bar{Z}\Phi_i)\Phi_i^T \sigma^{-2} + Z_i W S^{-1}\right)\right]\Sigma_{\theta_i} \tag{133}$$

$$\Sigma_{Z_i} \leftarrow \mathbb{E}_{\mathcal{Q}_{/Z_i}}\left[W S^{-1} W^T + I\right]^{-1} \tag{134}$$

$$\mu_{Z_i} \leftarrow \mathbb{E}_{\mathcal{Q}_{/Z_i}}\left[\theta_i S^{-1} W^T\right]\Sigma_{Z_i} \tag{135}$$

$$a_{\varsigma_k} \leftarrow a_0 + \frac{P}{2} \tag{136}$$

$$b_{\varsigma_k} \leftarrow \mathbb{E}_{\mathcal{Q}_{/\varsigma_k}}\left[b_0 + \frac{1}{2}\sum_i (\theta_{ik} - Z_i W_{\cdot k})^2 \beta_k\right] \tag{137}$$

$$\Sigma_{W_{\cdot k}} \leftarrow \mathbb{E}_{\mathcal{Q}_{/W_{\cdot k}}}\left[\varsigma_k^{-2}\beta_k \sum_i Z_i^T Z_i + \beta_k \,\mathrm{diag}(\alpha)\right]^{-1} \tag{138}$$

$$\mu_{W_{\cdot k}} \leftarrow \mathbb{E}_{\mathcal{Q}_{/W_{\cdot k}}}\left[\varsigma_k^{-2}\beta_k \sum_i (\theta_{ik} Z_i)\right]\Sigma_{W_{\cdot k}} \tag{139}$$

$$a_{\beta_k} \leftarrow a_0 + \frac{1 + K + P}{2} \tag{140}$$

$$b_{\beta_k} \leftarrow \mathbb{E}_{\mathcal{Q}_{/\beta_k}}\left[b_0 + \frac{1}{2}\left[\bar{Z}_k^2 \eta + \sum_j (W_{jk}^2 \alpha_j) + \sum_i (\theta_{ik} - Z_i W_{\cdot k})^2 \varsigma_k^{-2}\right]\right] \tag{141}$$

$$\Sigma_{\bar{Z}} \leftarrow \mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\sum_i (\Phi_i \Phi_i^T) + \eta \,\mathrm{diag}(\beta)\right]^{-1} \tag{142}$$

$$\mu_{\bar{Z}} \leftarrow \mathbb{E}_{\mathcal{Q}_{/\bar{Z}}}\left[\sigma^{-2}\sum_i \left[(Y_i - \theta_i \Phi_i)\Phi_i^T\right]\right]\Sigma_{\bar{Z}} \tag{143}$$

$$a_{\sigma^{-2}} \leftarrow a_0 + \frac{1}{2}\sum_i N_i \tag{144}$$

$$b_{\sigma^{-2}} \leftarrow \mathbb{E}_{\mathcal{Q}_{/\sigma}}\left[b_0 + \frac{1}{2}\sum_i ||Y_i - (\bar{Z} + \theta_i)\Phi_i||_2^2\right] \tag{145}$$

Notably, the columns of W becomes conditionally independent with the introduction of the slack variable θ, akin to the strategy described in [34,62]. Then the surrogate posterior of W factorizes over the columns, thereby requiring calculating the covariance of each column separately instead of the entire W at once. Thus, the computational complexity is significantly reduced. This factorization is

 doi:10.20944/preprints202503.0658.v1

introduced on top of the existing factorizations, thus the low-dimensional optimization strategy of BSFDA also applies to BSFDA$^{\text{Fast}}$.

11. Fast Initialization

In order to efficiently obtain a good initialization for the unknowns to be estimated, e.g. $Z, \bar{Z}, \beta$ and σ, we approximate the model so that we can adopt a fast strategy maximizing marginal likelihood using direct differentiation that is similar to [73]. This initial β serves to select the $K^{(a)}$ basis functions to start with.

We introduce $\tilde{Z}$ for easier marginalization:

$$Y_i = \tilde{Z}_i \Phi_i + E_i \tag{146}$$

$$\tilde{Z}_{ik} = \frac{Z_{ik}}{\sqrt{\beta_k}} + \bar{Z}_k \sim \mathcal{N}(\bar{Z}_k, \beta_k^{-1}) \tag{147}$$

$$\bar{Z}_k \sim \mathcal{N}(0, \beta_k^{-1}) \tag{148}$$

$$\beta_k \sim \Gamma(\beta_k | a_0, b_0), \sigma^{-2} \sim \Gamma(\sigma^{-2} | a_0, b_0) \tag{149}$$

$$E_i \sim \mathcal{N}(0, \sigma^2 I) \tag{150}$$

The approximated probabilistic graphical model is shown in Figure 11.

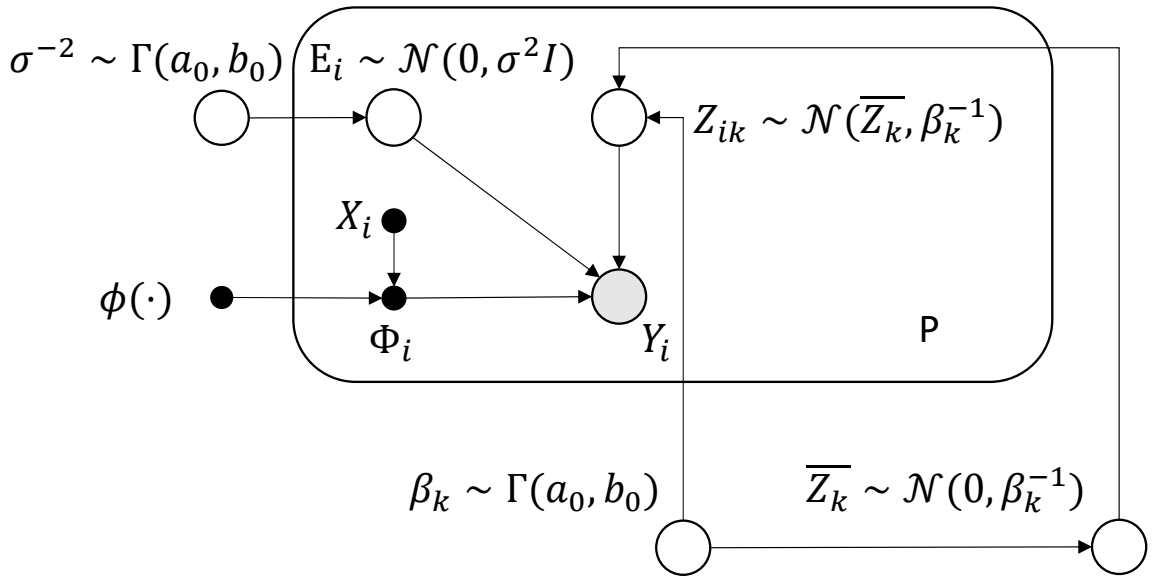


Figure 11. Probabilistic graphical model for the simplified model.

11.1. Maximum Likelihood Estimation

We apply maximum likelihood estimation for point estimates of $\bar{Z}, \beta, \sigma$.

$$\bar{Z}^*, \beta^*, \sigma^* \leftarrow \arg\min_{\bar{Z}, \beta, \sigma} \mathcal{P}, \tag{151}$$

where $\mathcal{P} = -\ln \Pr[Y | \bar{Z}, \beta, \sigma]$. Conditioned on these estimates, we can calculate the expectation of Z.

Optimization of $\beta, \bar{Z}$

 doi:10.20944/preprints202503.0658.v1

We set the differentiation to zero, i.e., $\frac{\partial \mathcal{P}}{\partial \beta_k} = 0$, and get:

$$\beta_k \leftarrow \begin{cases} \theta_k, \text{if } \theta_k > 0 \\ \infty, \text{otherwise} \end{cases} \tag{152}$$

where

$$\theta_k = \left(\frac{\sum_{i=1}^{P} s_{ik}^2}{\sum_{i=1}^{P} (q_{ik}^2 - s_{ik})} \right) \tag{153}$$

$$q_{ik} = \Phi_{ik} \mathcal{C}_{i/k}^{-1} (Y - \bar{Z}\Phi_i)^T \tag{154}$$

$$s_{ik} = \Phi_{ik} \mathcal{C}_{i/k}^{-1} \Phi_{ik}^T \tag{155}$$

$$\mathcal{C}_{i/k} = \mathcal{C}_i - \Phi_{ik}^T \beta_k^{-1} \Phi_{ik} \tag{156}$$

$$\mathcal{C}_i = \Phi_i^T \operatorname{diag}(\beta^{-1}) \Phi_i + \sigma^2 I = \sum_{k=1}^{K} \Phi_{ik}^T \beta_k^{-1} \Phi_{ik} + \sigma^2 I \tag{157}$$

We differentiate $\mathcal{P}$ with respect to $\bar{Z}$ and zero the derivative, i.e., $\frac{\partial \mathcal{P}}{\bar{Z}} = 0$, to get:

$$\bar{Z} \leftarrow \sum_{i=1}^{P} \left(Y_i \mathcal{C}_i^{-1} \Phi_i^T \right) \left(\sum_{i=1}^{P} (\Phi_i \mathcal{C}_i^{-1} \Phi_i^T) \right)^{-1} \tag{158}$$

We approximate Equation (158) by $\bar{Z}_A \leftarrow \sum_{i=1}^{P} \left(Y_i \mathcal{C}_i^{-1} \Phi_{iA}^T \right) \left(\sum_{i=1}^{P} (\Phi_{iA} \mathcal{C}_i^{-1} \Phi_{iA}^T) \right)^{-1}$ and $\bar{Z}_B \leftarrow 0$. This way we can apply the update with only the active basis functions.

Optimization of σ:

We use EM to optimize σ. In E step:

$$\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_i] \leftarrow \sigma^{-2} (Y_i - \bar{Z}\Phi_i) \Phi_i^T \mathcal{S}_i \tag{159}$$

$$\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_i^T \tilde{Z}_i] \leftarrow \mathcal{S}_i + \mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_i]^T \mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_i], \tag{160}$$

where $\mathcal{S}_i = (\Psi_i \sigma^{-2} + \operatorname{diag}(\beta))^{-1}$.

In M step

$$\begin{aligned} \sigma^{-2} &\leftarrow \frac{\sum_{i=1}^{P} \mathbb{E}_{\mathcal{Q}_{\tilde{Z}}} \left[||Y_i - (\tilde{Z}_i + \bar{Z})\Phi_i||_2^2 \right]}{\sum_{i=1}^{P} N_i} \\ &= \frac{\sum_{i=1}^{P} (Y_i - \bar{Z}\Phi_i)(Y_i - \bar{Z}\Phi_i - 2\Phi_i^T \mathbb{E}_{\mathcal{Q}_{\tilde{Z}}} [\tilde{Z}_i]^T)^T + \operatorname{Tr}(\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}} [\tilde{Z}_i^T \tilde{Z}_i] \Psi_i)}{\sum_{i=1}^{P} N_i} \end{aligned} \tag{161}$$

The optimization iterates between the E-step Equations (159) and (160), and the M-step Equation (161).

In practice, we need only $\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_{iA}], \mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_{iA}^T \tilde{Z}_{iA}], \mathcal{S}_{iA}$, and they can be calculated using the $K^{(a)}$ active basis functions. Thus, similar to [73], all the computations can be operated with only the active basis functions and thus it computationally efficient. This is described in Algorithm 3.

$$\mathcal{P} = -\ln \Pr[Y|\bar{Z}, \beta, \sigma] = -\sum_{i=1}^{P} \ln \Pr[Y_i|\bar{Z}, \beta, \sigma] = \sum_{i=1}^{P} \mathcal{P}_i \tag{162}$$

$$\mathcal{P}_i = \int \Pr[Y_i|\tilde{Z}_i, \bar{Z}, \beta, \sigma] \Pr[\tilde{Z}_i|\bar{Z}, \beta] d\tilde{Z}_i = \mathbb{E}_{\tilde{Z}_i \sim \mathcal{N}(\bar{Z}, \beta)} [\Pr[Y_i|\tilde{Z}_i, \sigma]] = \mathcal{N}(Y_i|\bar{Z}\Phi_i, \mathcal{C}_i) \tag{163}$$

$$\Pr[Y_i|\tilde{Z}_i, \bar{Z}, \beta, \sigma] = \mathcal{N}(Y_i|(\tilde{Z}_i + \bar{Z})\Phi_i, \sigma^2 I) \tag{164}$$

 doi:10.20944/preprints202503.0658.v1

Algorithm 3 Multi-sample relevance vector machine

while $\mathcal{P}$ is not converged **do**
  $k \leftarrow$ a random number that satisfies $\text{CosSim}(\phi_k, \phi_A) \leq \tau_{\text{sim}}$ ▷ $\mathcal{O}(K^{(\text{a})^3})$
  $s_{ik} \leftarrow \Phi_{ik}\mathcal{C}_{i/_k}^{-1}\Phi_{ik}^T, \forall i$ ▷ Sparsity factor. $\mathcal{O}\left(P\max(K^{(\text{a})^3}, \max_i(N_i)^2)\right)$
  $q_{ik} \leftarrow \Phi_{ik}\mathcal{C}_{i/_k}^{-1}(Y - \bar{Z}_A\Phi_{iA})^T, \forall i$ ▷ Quality factor. $\mathcal{O}\left(P\max_i(N_i)\max(K^{(\text{a})}, \max_i(N_i))\right)$
  $\theta_k \leftarrow \left(\frac{\sum_{i=1}^P s_{ik}^2}{\sum_{i=1}^P (q_{ik}^2 - s_{ik})}\right)$
  if $\theta > 0$ **then**
    $\beta_k \leftarrow \theta_k$ ▷ Precision is finite
  else
    $\beta_k \leftarrow \infty$ ▷ Precision is infinite and the dimension is removed
  end if
  $\Phi_{iA} \leftarrow$ All Φ_{ik} that has $\beta_k < \infty, \forall i$
  $\mathcal{C}_i = \sum_{\beta_k<\infty} \Phi_{ik}^T\beta_k^{-1}\Phi_{ik} + \sigma^2 I, \forall i$
  $\bar{Z}_A \leftarrow \sum_{i=1}^P \left(Y_i\mathcal{C}_i^{-1}\Phi_{iA}^T\right)\left(\sum_{i=1}^P(\Phi_{iA}\mathcal{C}_i^{-1}\Phi_{iA}^T)\right)^{-1}$ ▷ $\mathcal{O}\left(PK^{(\text{a})}\max\left(K^{(\text{a})}, \max_i(N_i)\right)^2\right)$
  $\mathcal{S}_{iA} \leftarrow (\Phi_{iA}\Phi_{iA}^T\sigma^{-2} + \text{diag}(\beta_A))^{-1}, \forall i$
  $\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_{iA}] \leftarrow \sigma^{-2}(Y_i - \bar{Z}_A\Phi_{iA})\Phi_{iA}^T\mathcal{S}_{iA}, \forall i$ ▷ $\mathcal{O}(PK^{(\text{a})^2}\max_i(N_i))$
  $\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_{iA}^T\tilde{Z}_{iA}] \leftarrow \mathcal{S}_{iA} + \mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_{iA}]^T\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_{iA}], \forall i$
  $\sigma \leftarrow \frac{\sum_{i=1}^P (Y_i - \bar{Z}_A\Phi_{iA})(Y_i - \bar{Z}_A\Phi_{iA} - 2\Phi_{iA}^T\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_{iA}]^T)^T + \text{Tr}(\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_{iA}^T\tilde{Z}_{iA}]\Phi_{iA}\Phi_{iA}^T)}{\sum_{i=1}^P N_i}$
end while

We apply Sylvester's determinant theorem to Equation (157) and get

$$|\mathcal{C}_i| = |\mathcal{C}_{i_{/_k}}||I + \beta_k^{-1}\Phi_{ik}^T\mathcal{C}_{i_{/_k}}^{-1}\Phi_{ik}| \tag{165}$$

We apply Woodbury matrix identity to Equation (157) and get

$$\mathcal{C}_i^{-1} = \mathcal{C}_{i_{/_k}}^{-1} - \mathcal{C}_{i_{/_k}}^{-1}\Phi_{ik}^T(\beta_k + \Phi_{ik}\mathcal{C}_{i_{/_k}}^{-1}\Phi_{ik}^T)^{-1}\Phi_{ik}\mathcal{C}_{i_{/_k}}^{-1} \tag{166}$$

We first expand $\mathcal{P}_i$

$$\begin{aligned}
\mathcal{P}_i &= \ln \Pr[Y_i|\bar{Z}, \sigma, \beta] \\
&= -\frac{1}{2}\sum_i \ln|2\pi\mathcal{C}_i| + (Y_i - \bar{Z}\Phi_i)\mathcal{C}_i^{-1}(Y_i - \bar{Z}\Phi_i)^T \\
&= -\frac{1}{2}(N_i\ln(2\pi) + \ln|\mathcal{C}_{i_{/_k}}| + \ln|I + \beta_k^{-1}\Phi_{ik}\mathcal{C}_{i_{/_k}}^{-1}\Phi_{ik}^T| + (Y_i - \bar{Z}\Phi_i)\mathcal{C}_i^{-1}(Y_i - \bar{Z}\Phi_i)^T \\
&- (\beta_k + \Phi_{ik}\mathcal{C}_{i_{/_k}}^{-1}\Phi_{ik}^T)^{-1}||\Phi_{ik}\mathcal{C}_{i_{/_k}}^{-1}(Y - \bar{Z}\Phi_i)^T||_2^2) \\
&= \mathcal{P}_{i_{/_k}} + \frac{1}{2}(\ln\beta_k - \ln|\beta_k + s_{ik}| + \frac{q_{ik}^2}{\beta_k + sik})
\end{aligned} \tag{167}$$

where we plug in Equations (165) and (166) and define q_{ik}, s_{ik} in a similar way to [73]. The sparsity factor s_{ik} can be seen to be a measure of the extent that the basis function ϕ_k overlaps those already present in the model under the measurements at index set X_i. The quality factor q_{ik} is a measure of the alignment with the error of the model at X_i with that basis function excluded. Because we are representing the mean functions using only the active basis functions, i.e., $\bar{Z}_k = 0$ when $\beta_k = \infty$, Equation (154) only uses the K active basis functions. Similarly, Equation (155) only uses the K active basis functions.

For computational efficiency, we can compute s_{ik}, q_{ik} using $S_{ik} = \Phi_{ik}\mathcal{C}_i^{-1}\Phi_{ik}^T, Q_{ik} = \Phi_{ik}\mathcal{C}_i^{-1}(Y - \bar{Z}\Phi_i)^T$ in a similar way to [73] as follows

 doi:10.20944/preprints202503.0658.v1

$$
\begin{aligned}
s_{ik} &= \Phi_{ik}\mathcal{C}_{i/k}^{-1}\Phi_{ik}^T = S_{ik} + \Phi_{ik}\mathcal{C}_{i/k}^{-1}\Phi_{ik}^T(\beta_k + \Phi_{ik}\mathcal{C}_{i/k}^{-1}\Phi_{ik}^T)^{-1}\Phi_{ik}\mathcal{C}_{i/k}^{-1}\Phi_{ik}^T \\
&= S_{ik} + s_{ik}(\beta_k + s_{ik})^{-1}s_{ik} \rightleftharpoons s_{ik} = \frac{\beta_k + s_{ik}}{\beta_k}S_{ik} \qquad (168) \\
\Rightarrow s_{ik} &\leftarrow (1 - \frac{1}{\beta_k}S_{ik})^{-1}S_{ik} = \frac{\beta_k S_{ik}}{\beta_k - S_{ik}} \qquad (169) \\
q_{ik} &= \Phi_{ik}\mathcal{C}_{i/k}^{-1}(Y - \bar{Z}\Phi_i)^T \\
&= Q_{ik} + \Phi_{ik}\mathcal{C}_{i/k}^{-1}\Phi_{ik}^T(\beta_k + \Phi_{ik}\mathcal{C}_{i/k}^{-1}\Phi_{ik}^T)^{-1}\Phi_{ik}\mathcal{C}_{i/k}^{-1}(Y - \bar{Z}\Phi_i)^T \\
&= Q_{ik} + s_{ik}(\beta_k + s_{ik})^{-1}q_{ik} \qquad (170) \\
\Rightarrow q_{ik} &\leftarrow \frac{\beta_k + s_{ik}}{\beta_k}Q_{ik} = \frac{\beta_k Q_{ik}}{\beta_k - S_{ik}} \qquad (171)
\end{aligned}
$$

11.2. Optimization of $\beta, \bar{Z}$

Derivation of Equation (152):

We differentiate $\mathcal{P}$ with respect to β_k

$$
\begin{aligned}
\frac{\partial \mathcal{P}}{\partial \beta_k} &= \sum_{i=1}^{P}\frac{1}{2}\left(\beta_k^{-1} - |\beta_k + s_{ik}|^{-1} - q_{ik}^2(\beta_k + s_{ik})^2\right) \\
&= \frac{1}{2}\beta_k^{-1}\sum_{i=1}^{P}\left((\beta_k + s_{ik})^{-2}(\beta_k(s_{ik} - q_{ik}^2) + s_{ik}^2)\right) \qquad (172)
\end{aligned}
$$

We further adopt the approximation $s_{1k} \approx s_{2k} \approx \ldots \approx s_{Pk}$. Because s_{ik} is a discrete measure of the overlapping between the basis functions, it should remain invariant with respect to different sampling grid X_i given the number of measurements is adequate and similar. Alternatively, the Expectation maximization scheme can also be applied and is guaranteed to increase the likelihood $\mathcal{P}$ in each iteration until convergence. However, we opt for this gradient descent with approximations for its advantage in speed to obtain a reasonable initialization. This way we set the approximated differentiation to zero

$$
\begin{aligned}
\frac{\partial \mathcal{P}}{\partial \beta_k} &\approx \frac{1}{2}\beta_k^{-1}(\beta_k + s_{1k})^{-2}\sum_{i=1}^{P}\left((\beta_k(s_{ik} - q_{ik}^2) + s_{ik}^2)\right) = 0 \qquad (173) \\
\Rightarrow \beta_k &\leftarrow \theta_k = \left(\frac{\sum_{i=1}^{P} s_{ik}^2}{\sum_{i=1}^{P}(q_{ik}^2 - s_{ik})}\right) \qquad (174)
\end{aligned}
$$

Because β_k is a scale parameter, we need $\beta_k > 0$. Consequently, the optimal value for β_k to maximize $\mathcal{P}$ dependents on the sign of θ_k. When $\theta_k > 0$, the maximum of $\mathcal{P}$ is achieved at $\beta_k = \theta_k$.

On the other hand when $\theta_k \leq 0$, $\mathcal{P}$ is monotonically increasing with respect to β_k and therefore we should have $\beta_k \leftarrow \infty$ in order to maximize $\mathcal{P}$.

More intuitively, Equation (174) can be regarded as a weighted summation of the estimation of β_k using each individual sample function and it automatically assigns more weights to those with more measurements. Therefore, this optimization strategy is supposed to provide reasonable estimates even when the sampled functions have different numbers of measurements.

Derivation of Equation (158):

 doi:10.20944/preprints202503.0658.v1

We differentiate $\mathcal{P}$ with respect to $\bar{Z}$ and zero the derivative to get

$$\frac{\partial \mathcal{P}}{\bar{Z}} = -\frac{1}{2}\sum_{i=1}^{P}\left(-2Y_i\mathcal{C}_i\Phi_i^T + 2\bar{Z}\Phi_i\mathcal{C}_i^{-1}\Phi_T\right) = 0 \tag{175}$$

$$\Rightarrow \bar{Z} \leftarrow \sum_{i=1}^{P}\left(Y_i\mathcal{C}_i^{-1}\Phi_i^T\right)\left(\sum_{i=1}^{P}(\Phi_i\mathcal{C}_i^{-1}\Phi_i^T)\right)^{-1} \tag{176}$$

11.3. Optimization of σ

Derivation of Equations (159) and (160):

We use the Expectation maximization strategy with latent variables $\tilde{Z}_i$. It is similar to that used in [52]. It introduces a surrogate function, the log likelihood for the complete data $\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\mathcal{P}_{\tilde{Z}}]$, that is easier to optimize and in theory the process ultimately maximizes $\mathcal{P}$.

For the E-step, we calculate the posterior of $\tilde{Z}_i$.

$$\ln \Pr[\tilde{Z}_i|Y_i,\bar{Z},\sigma,\beta] = \ln\frac{\Pr[Y_i|\tilde{Z}_i,\bar{Z},\sigma,\beta]\Pr[\tilde{Z}_i|\beta]}{\Pr[Y_i|\bar{Z},\sigma,\beta]} \tag{177}$$

$$\propto -\frac{1}{2}\left(\tilde{Z}_i(\Psi_i\sigma^{-2} + \mathrm{diag}(\beta))\tilde{Z}_i^T - 2\sigma^{-2}(Y_I - \bar{Z}\Phi_i)\Phi_i^T\tilde{Z}_i^T\right) \tag{178}$$

Therefore,

$$\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_i] \leftarrow \sigma^{-2}(Y_i - \bar{Z}\Phi_i)\Phi_i^T\mathcal{S}_i \tag{179}$$

$$\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_i^T\tilde{Z}_i] \leftarrow \mathcal{S}_i + \mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_i]^T\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\tilde{Z}_i] \tag{180}$$

where

$$\mathcal{S}_i = (\Psi_i\sigma^{-2} + \mathrm{diag}(\beta))^{-1} \tag{181}$$

Derivation of Equation (161):

In M-step, we need to maximum $\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\mathcal{P}_{\tilde{Z}}]$ conditioned on $\mathcal{Q}_{\tilde{Z}}$ with respect to σ^{-2},

$$\begin{aligned}\mathcal{P}_{\tilde{Z}} &= \sum_{i=1}^{P}\ln\Pr[Y_i,\tilde{Z}_i|\bar{Z},\sigma,\beta] = \sum_{i=1}^{P}\ln(\Pr[Y_i|\tilde{Z}_i,\bar{Z},\sigma]\Pr[\tilde{Z}_i|\beta])\\ &= -\frac{1}{2}\sum_{i=1}^{P}\Bigg(N_i\ln(2\pi\sigma^{-2}) + \sigma^{-2}||Y_i - (\tilde{Z}_i+\bar{Z})\Phi_i||_2^2 + \\ &\sum_{k=1}^{K}\ln(2\pi\beta_k^{-1}) + \mathrm{Tr}(\tilde{Z}_i\,\mathrm{diag}(\beta)\tilde{Z}_i^T)\Bigg)\end{aligned} \tag{182}$$

We differentiate $\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\mathcal{P}_{\tilde{Z}}]$ with respect to σ^{-2} and set to 0

$$\frac{\partial\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}[\mathcal{P}_{\tilde{Z}}]}{\partial\sigma^{-2}} = \mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}\left[-\frac{1}{2}\sum_{i=1}^{P}\left(N_i\sigma^{-2} - \sigma^{-4}||Y_i - (\tilde{Z}_i+\bar{Z})\Phi_i||_2^2\right)\right] = 0 \tag{183}$$

$$\begin{aligned}\Rightarrow \sigma^{-2} &\leftarrow \frac{\sum_{i=1}^{P}\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}\left[||Y_i - (\tilde{Z}_i+\bar{Z})\Phi_i||_2^2\right]}{\sum_{i=1}^{P}N_i}\\ &= \frac{\sum_{i=1}^{P}(Y_i - \bar{Z}\Phi_i)(Y_i - \bar{Z}\Phi_i - 2\Phi_i^T\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}\left[\tilde{Z}_i\right]^T)^T + \mathrm{Tr}(\mathbb{E}_{\mathcal{Q}_{\tilde{Z}}}\left[\tilde{Z}_i^T\tilde{Z}_i\right]\Psi_i)}{\sum_{i=1}^{P}N_i}\end{aligned} \tag{184}$$

 doi:10.20944/preprints202503.0658.v1

12. Experiments

12.1. Benchmark Simulation

Figure 12 presents the application of the proposed BSFDA to the simulation benchmark (Scenario 1) outlined in [37]. While prior analyses have utilized this benchmark, the current experimental configuration is specifically adapted to highlight the method's capacity for uncertainty quantification. The experimental design consists of 20 functional observations, each sampled at either 3 points (with a 20% probability) or 10 points (with an 80% probability), determined via random assignment. The number of sampled functions is decreased from 200 to 20 to underscore the effect and estimation of uncertainties. The actual white noise standard deviation is 0.4472, while the estimated standard deviation is 0.4839. The component number is also correctly estimated as 3. The figure depicts the true underlying function, the discrete observational data, and the corresponding functional estimates, accompanied by their respective 95% truncated uncertainty intervals.

Notably, the uncertainty associated with sparsely sampled functions exhibits substantial inflation in regions devoid of observations. In contrast, in sampled regions, the uncertainty aligns closely with that of densely sampled functions, approximating twice the standard deviation of the white noise. Additionally, the uncertainty bounds for the estimated mean function are presented, demonstrating reduced variability relative to individual function estimates.

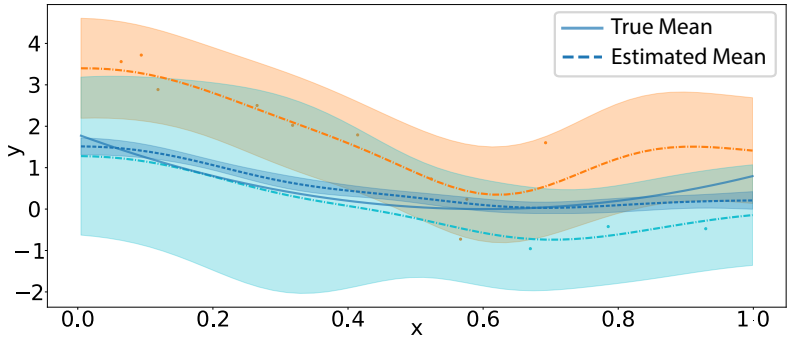


Figure 12. Application of the proposed BSFDA to the simulation benchmark from [37], illustrating the true mean function (blue), observed measurements from two functions sampled at different densities (light blue for sparse, orange for dense), and the corresponding functional estimates with 95% truncated uncertainty intervals.

Table 9. Distributions of the estimated component number $\hat{r}$ for Scenario 1 (r=3).

N_i	$\hat{r}$	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	$\text{AIC}^{2022}_{\text{PACE}}$	$\text{BIC}^{2022}_{\text{PACE}}$	fpca	BSFDA	$\text{BSFDA}^{\text{Fast}}$
5	≤ 1	0.000	0.000	0.155	0.005	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.008	0.405	0.335	0.565	0.215	0.000	0.000	0.000	0.000	0.985
	=3	0.000	0.580	0.380	0.410	0.735	0.650	0.880	0.645	**0.995**	0.015
	=4	0.121	0.010	0.115	0.010	0.045	0.335	0.120	0.235	0.005	0.000
	≥ 5	0.870	0.005	0.015	0.010	0.005	0.015	0.000	0.120	0.000	0.000
10	≤ 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.000	0.005	0.040	0.040	0.005	0.000	0.000	0.000	0.000	0.075
	=3	0.000	0.980	0.670	0.955	0.985	0.880	0.920	0.645	**1.000**	0.910
	=4	0.000	0.015	0.255	0.000	0.010	0.120	0.080	0.235	0.000	0.015
	≥ 5	1.000	0.000	0.035	0.005	0.000	0.000	0.000	0.120	0.000	0.000
50	≤ 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=3	0.000	**1.000**	0.830	**1.000**	**1.000**	**1.000**	**1.000**	0.890	0.980	0.945
	=4	0.000	0.000	0.150	0.000	0.000	0.000	0.000	0.060	0.020	0.050
	≥ 5	1.000	0.000	0.020	0.000	0.000	0.000	0.000	0.050	0.000	0.005

 doi:10.20944/preprints202503.0658.v1

Table 10. Distributions of the estimated component number $\hat{r}$ for Scenario 2 (r=3).

N_i	$\hat{r}$	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	AIC^{2022}_{PACE}	BIC^{2022}_{PACE}	fpca	BSFDA	$BSFDA^{Fast}$
5	≤ 1	0.000	0.000	0.230	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.000	0.205	0.395	0.000	0.140	0.050	0.075	0.000	0.000	0.960
	=3	0.005	0.630	0.245	0.375	0.605	0.570	0.620	0.475	**1.000**	0.040
	=4	0.125	0.155	0.110	0.440	0.210	0.345	0.275	0.350	0.000	0.000
	≥ 5	0.870	0.010	0.020	0.185	0.045	0.035	0.030	0.175	0.000	0.000
10	≤ 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.000	0.000	0.170	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=3	0.000	0.710	0.665	0.570	0.805	0.825	0.850	0.640	**1.000**	0.995
	=4	0.005	0.260	0.135	0.355	0.185	0.175	0.150	0.235	0.000	0.005
	≥ 5	0.995	0.030	0.030	0.075	0.010	0.000	0.000	0.125	0.000	0.000
50	≤ 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=3	0.000	0.630	0.795	0.955	0.945	**1.000**	**1.000**	0.950	**1.000**	0.950
	=4	0.000	0.320	0.185	0.045	0.055	0.000	0.000	0.020	0.000	0.050
	≥ 5	1.000	0.050	0.020	0.000	0.000	0.000	0.000	0.030	0.000	0.000

Table 11. Distributions of the estimated component number $\hat{r}$ for Scenario 3 (r=3).

N_i	$\hat{r}$	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	AIC^{2022}_{PACE}	BIC^{2022}_{PACE}	fpca	BSFDA	$BSFDA^{Fast}$
5	≤ 1	0.000	0.000	0.335	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.025	0.035	0.260	0.220	0.005	0.000	0.005	0.000	0.000	0.025
	=3	0.005	0.720	0.325	0.640	0.590	0.320	0.400	0.450	**0.995**	0.945
	=4	0.130	0.170	0.080	0.075	0.280	0.640	0.565	0.360	0.005	0.030
	≥ 5	0.840	0.075	0.000	0.065	0.125	0.030	0.030	0.190	0.000	0.000
10	≤ 1	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.015	0.000	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=3	0.000	0.580	0.770	0.965	0.665	0.740	0.755	0.440	**0.995**	1.000
	=4	0.000	0.400	0.145	0.030	0.320	0.260	0.245	0.380	0.005	0.000
	≥ 5	0.985	0.020	0.045	0.005	0.015	0.000	0.000	0.180	0.000	0.000
50	≤ 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.000
	=3	0.000	1.000	0.775	1.000	1.000	**1.000**	**1.000**	0.765	0.980	0.920
	=4	0.000	0.000	0.200	0.000	0.000	0.000	0.000	0.110	0.005	0.050
	≥ 5	1.000	0.000	0.025	0.000	0.000	0.000	0.000	0.125	0.000	0.030

Table 12. Distributions of the estimated component number $\hat{r}$ for Scenario 4 (r=3).

N_i	$\hat{r}$	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	AIC^{2022}_{PACE}	BIC^{2022}_{PACE}	fpca	BSFDA	$BSFDA^{Fast}$
5	≤ 1	0.000	0.000	0.315	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.015	0.020	0.180	0.160	0.015	0.000	0.000	0.000	0.000	0.000
	=3	0.015	0.710	0.410	0.640	0.560	0.515	0.575	0.370	**1.000**	0.975
	=4	0.145	0.185	0.070	0.095	0.260	0.450	0.390	0.515	0.000	0.025
	≥ 5	0.825	0.085	0.025	0.105	0.165	0.035	0.035	0.115	0.000	0.000
10	≤ 1	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=3	0.000	0.830	0.775	0.920	0.900	0.750	0.760	0.350	**0.995**	0.990
	=4	0.000	0.150	0.190	0.045	0.085	0.250	0.240	0.380	0.005	0.010
	≥ 5	1.000	0.020	0.020	0.035	0.015	0.000	0.000	0.270	0.000	0.000
50	≤ 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.000
	=3	0.000	0.945	0.835	**1.000**	**1.000**	**1.000**	**1.000**	0.730	0.950	0.935
	=4	0.000	0.055	0.140	0.000	0.000	0.000	0.000	0.160	0.040	0.055
	≥ 5	1.000	0.000	0.025	0.000	0.000	0.000	0.000	0.110	0.000	0.010

 doi:10.20944/preprints202503.0658.v1

Table 13. Distributions of the estimated component number $\hat{r}$ for Scenario 5 (r=6).

N_i	$\hat{r}$	AIC_{PACE}	AIC	BIC	PC_{p1}	IC_{p1}	AIC^{2022}_{PACE}	BIC^{2022}_{PACE}	fpca	BSFDA	$BSFDA^{Fast}$
5	≤ 4	0.005	0.165	0.835	0.580	0.060	0.000	0.000	0.010	0.000	0.060
	=5	0.005	0.330	0.020	0.345	0.335	0.575	0.590	0.010	0.075	0.515
	=6	0.705	0.470	0.090	0.070	0.545	0.425	0.410	0.855	**0.925**	0.160
	=7	0.245	0.035	0.050	0.005	0.060	0.000	0.000	0.115	0.000	0.160
	≥ 8	0.040	0.000	0.005	0.000	0.000	0.000	0.000	0.010	0.000	0.105
10	≤ 4	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=5	0.000	0.000	0.030	0.145	0.000	0.425	0.425	0.000	0.000	0.000
	=6	0.065	0.570	0.525	0.775	0.705	0.575	0.575	0.500	**1.000**	0.930
	=7	0.475	0.280	0.165	0.020	0.185	0.000	0.000	0.405	0.000	0.035
	≥ 8	0.455	0.150	0.030	0.060	0.110	0.000	0.000	0.095	0.000	0.035
50	≤ 4	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	=5	0.065	0.000	0.000	0.000	0.000	0.130	0.130	0.005	0.000	0.000
	=6	0.000	0.260	0.590	0.980	0.965	0.870	0.770	0.695	**0.995**	0.925
	=7	0.000	0.405	0.325	0.010	0.035	0.000	0.000	0.250	0.005	0.045
	≥ 8	0.935	0.335	0.080	0.010	0.000	0.000	0.000	0.050	0.000	0.030

12.1.1. Performance of LFRM

To compare the latent factor regression model (LFRM) [34] as a dimension reduction model to ours, Bayesian scalable functional data analysis (BSFDA), we set the covariates in LFRM to zero, thus assigning standard Gaussian priors to the latent variables, analogous to our approach. We followed the simulation benchmark in [37] for selecting the number of components, focusing on Scenario 1 with 50 measurements per function (the densest data). Because LFRM does not estimate a mean function, we omitted the mean from the simulation run here.

The following hyperparameters of LFRM need to be determined:

- Gamma prior for white noise and correlated noise
- Length-scale
- Number of basis functions
- Number of iterations

LFRM, with its default white-noise prior, correctly identified the white-noise variance (true value 0.2) in all tests. We thus retained that default. We tested different Gamma priors for correlated noise: the default prior, a noninformative-like (vagor) prior (same mean but 100 times the variance), and a low noise prior (same variance but 100 times the mean). We maintained the number of locations for basis functions at 10, which is the default setting. For length-scale in LFRM, we first used the best estimate from our cross-validation (CV). We then tried all 10 CV-selected length scales, producing 100 basis functions in total. However, this required substantial time, so we performed only two repeated runs for that setting. We kept LFRM's default of 5000 burn-in iterations (25,000 total) with thinning at intervals of 5, verifying convergence through trace plots in line with [34]. Meanwhile, BSFDA was run 200 times as in Section 5, LFRM (10 length scales) 2 times, and all other settings 10 times.

Across repeated trials, LFRM consistently overestimated the true number of components (which is 3). Specifically:

- Standard LFRM estimated 10–14 components.
- LFRM with 10 length-scales estimated 6–8 components.
- LFRM with a low correlated-noise prior estimated 8–15 components.
- LFRM with a noninformative-like correlated-noise prior estimated 10–14 components.

In contrast, our method BSFDA produced a clear gap in the distribution of the precision parameters, effectively separating effective dimensions from redundant ones.

Several factors may explain LFRM's performance:

- **Correlated noise interference:** The correlated noise can obscure the true signal.

 doi:10.20944/preprints202503.0658.v1

- **Prior specification:** LFRM's precision parameter prior are potentially less noninformative and not as sparse as those sparse Bayesian learning priors [52] in BSFDA.
- **Element-wise vs. Column-wise Precision:** The element-wise precision parameters in LFRM might compensate in a way that reduces overall sparsity.

12.2. Variational Inference v.s. MCMC

We conducted experiments using both Gibbs sampling (MCMC) and mean-field approximation (Variational Inference) for the Bayesian PCA simulation [45,62] under varying noise levels. In our experiments, "satisfactory estimation" is defined as the point when the 4th smallest precision (i.e., the inverse of variance) is at least 100 times smaller than the 5th–indicating that the four true signal dimensions (with variances [5, 4, 3, 2]) have been correctly identified. For computational tractability, we capped VI at 200,000 iterations (approximately 200 seconds) and MCMC sampling at 20,000 iterations (about 20 minutes), with a burn-in period of 200 iterations and thinning set to 10.

Figure 13 illustrates the runtime for VI and MCMC to identify the correct components. Our key findings are as follows:

1. When the noise level is close to the signal, neither MCMC or VI found the true dimension in the limited iterations (probably never will), because the data is heavily polluted.
2. As the noise level decreases toward zero, the number of iterations (and runtime) required for satisfactory estimation increases dramatically; VI begins to fail around a noise level of 1×10^{-4}, and MCMC sampling around 1×10^{-3}, within the set time constraints.
3. Across the 10 noise levels (about 3×10^{-3} to 2×10^{-1}) where both successfully identified the correct dimensionality, VI consistently completes much faster than MCMC sampling. VI is approximately 20 times faster. 85.57 ± 50.24 in average, in the range of 32.46 to 189.12.

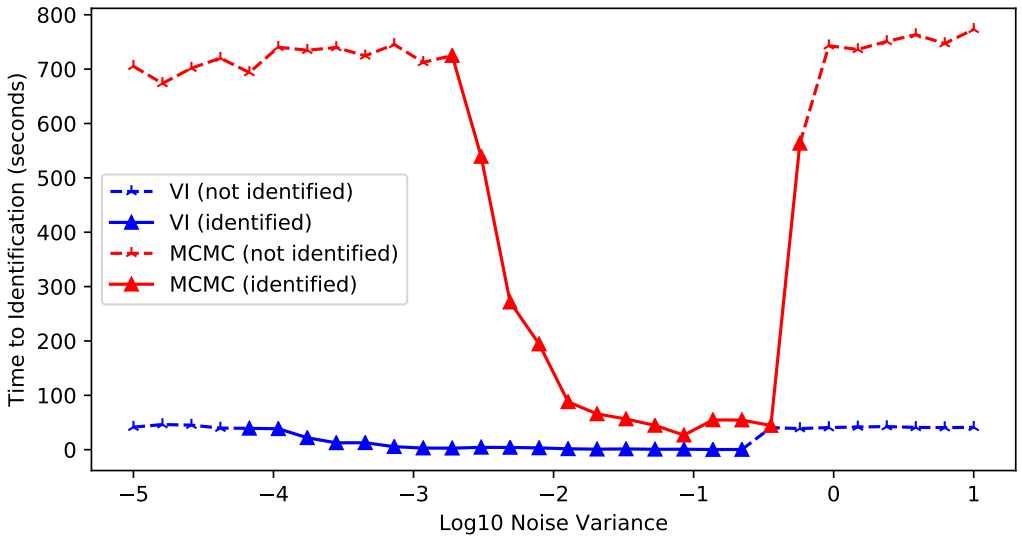


Figure 13. Time for variational inference and MCMC to identify the correct components in Bayesian PCA.

These results indicate that both MCMC sampling and VI become slower as noise decreases due to strong dependencies in the posterior. We hypothesize it is because both MCMC and VI suffer from the dependency introduced by low noise, which is a known long-standing issue with ongoing research methods, e.g., structured VI [72], or blocked/collapsed Gibbs sampler [74]. But both MCMC and VI work well provided there are sufficient iterations. This behavior suggests that the dependency induced by very low noise levels creates an optimization challenge rather than a fundamental modeling issue.

In summary: (1) VI is significantly faster than MCMC, (2) both methods slow down as the noise level decreases, and (3) both fail to recover the correct components when the noise is excessively high.

 doi:10.20944/preprints202503.0658.v1

References

1. Wang, J.L.; Chiou, J.M.; Müller, H.G. Functional Data Analysis. *Annual Review of Statistics and Its Application* **2016**, *3*, 257–295. https://doi.org/10.1146/annurev-statistics-041715-033624.
2. Ramsay, J.O.; Silverman, B.W. *Applied functional data analysis: methods and case studies*; Springer, 2002.
3. Rice, J.A.; Silverman, B.W. Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **1991**, *53*, 233–243. https://doi.org/10.1111/j.2517-6161.1991.tb01821.x.
4. Górecki, T.; Krzyśko, M.; Waszak, Ł.; Wołyński, W. Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers* **2018**, *59*, 153–182. https://doi.org/10.1007/s00362-016-0757-8.
5. Aneiros, G.; Cao, R.; Fraiman, R.; Genest, C.; Vieu, P. Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis* **2019**, *170*, 3–9. https://doi.org/10.1016/j.jmva.2018.11.007.
6. Li, Y.; Qiu, Y.; Xu, Y. From multivariate to functional data analysis: Fundamentals, recent developments, and emerging areas. *Journal of Multivariate Analysis* **2022**, *188*, 104806. https://doi.org/10.1016/j.jmva.2021.104806.
7. Happ, C.; Greven, S. Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association* **2018**, *113*, 649–659. https://doi.org/10.1080/01621459.2016.1273115.
8. Kowal, D.R.; Canale, A. Semiparametric Functional Factor Models with Bayesian Rank Selection. *Bayesian Analysis* **2023**, *18*. https://doi.org/10.1214/23-BA1410.
9. Paulon, G.; Müller, P.; Sarkar, A. Bayesian Semiparametric Hidden Markov Tensor Models for Time Varying Random Partitions with Local Variable Selection. *Bayesian Analysis* **2024**, *19*. https://doi.org/10.1214/23-BA1383.
10. Goldsmith, J.; Zipunnikov, V.; Schrack, J. Generalized Multilevel Function-on-Scalar Regression and Principal Component Analysis. *Biometrics* **2015**, *71*, 344–353. https://doi.org/10.1111/biom.12278.
11. Nguyen, X.; Gelfand, A.E. The Dirichlet labeling process for clustering functional data. *Statistica Sinica* **2011**, *21*, 1249–1289. https://doi.org/10.5705/ss.2008.285.
12. Petrone, S.; Guindani, M.; Gelfand, A.E. Hybrid Dirichlet Mixture Models for Functional Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2009**, *71*, 755–782. https://doi.org/10.1111/j.1467-9868.2009.00708.x.
13. Berrendero, J.; Justel, A.; Svarc, M. Principal components for multivariate functional data. *Computational Statistics & Data Analysis* **2011**, *55*, 2619–2634. https://doi.org/10.1016/j.csda.2011.03.011.
14. Suarez, A.J.; Ghosal, S. Bayesian Estimation of Principal Components for Functional Data. *Bayesian Analysis* **2017**, *12*. https://doi.org/10.1214/16-BA1003.
15. Yang, J.; Zhu, H.; Choi, T.; Cox, D.D. Smoothing and Mean–Covariance Estimation of Functional Data with a Bayesian Hierarchical Model. *Bayesian analysis* **2016/09**, *11*, 649–670. https://doi.org/10.1214/15-ba967.
16. Fox, E.B.; Dunson, D.B. Bayesian nonparametric covariance regression. *The Journal of Machine Learning Research* **2015**, *16*, 2501–2542.
17. Goldsmith, J.; Wand, M.P.; Crainiceanu, C. Functional regression via variational Bayes. *Electronic journal of statistics* **2011/01/01**, *5*, 572–602. https://doi.org/10.1214/11-ejs619.
18. Sun, T.Y.; Kowal, D.R. Ultra-Efficient MCMC for Bayesian Longitudinal Functional Data Analysis. *Journal of Computational and Graphical Statistics* **2024**, pp. 1–13. https://doi.org/10.1080/10618600.2024.2362227.
19. Rasmussen, C.E.; Williams, C.K.I. *Gaussian processes for machine learning*; Adaptive computation and machine learning, MIT Press: Cambridge, Mass, 2006; p. 248.
20. Yao, F.; Müller, H.G.; Wang, J.L. Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association* **2005**, *100*, 577–590. https://doi.org/10.1198/016214504000001745.
21. Di, C.Z.; Crainiceanu, C.M.; Caffo, B.S.; Punjabi, N.M. Multilevel functional principal component analysis. *The Annals of Applied Statistics* **2009**, *3*, 458–488. https://doi.org/10.1214/08-AOAS206.
22. Peng, J.; Paul, D. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics* **2009**, *18*, 995–1015.
23. Goldsmith, J.; Schwartz, J.E. Variable selection in the functional linear concurrent model. *Statistics in Medicine* **2017/06/30**, *36*, 2237–2250. https://doi.org/10.1002/sim.7254.

 doi:10.20944/preprints202503.0658.v1

24. Xiao, L.; Li, C.; Checkley, W.; Crainiceanu, C. Fast covariance estimation for sparse functional data. *Statistics and Computing* **2018**, *28*, 511–522. https://doi.org/10.1007/s11222-017-9744-8.
25. Chiou, J.M.; Yang, Y.F.; Chen, Y.T. Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica* **2014**. https://doi.org/10.5705/ss.2013.305.
26. Jacques, J.; Preda, C. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* **2014**, *71*, 92–106. https://doi.org/10.1016/j.csda.2012.12.004.
27. Yang, J.; Cox, D.D.; Lee, J.S.; Ren, P.; Choi, T. Efficient Bayesian Hierarchical Functional Data Analysis with Basis Function Approximations Using Gaussian–Wishart Processes. *Biometrics* **2017/12/01**, *73*, 1082–1091. https://doi.org/10.1111/biom.12705.
28. Trefethen, L.N. *Approximation theory and approximation practice, extended edition*; SIAM, 2019.
29. Bungartz, H.J.; Griebel, M. Sparse grids. *Acta numerica* **2004**, *13*, 147–269.
30. Zipunnikov, V.; Caffo, B.; Yousem, D.M.; Davatzikos, C.; Schwartz, B.S.; Crainiceanu, C. Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics* **2011**, *20*, 852–873. https://doi.org/10.1198/jcgs.2011.10122.
31. Yao, F.; Lei, E.; Wu, Y. Effective dimension reduction for sparse functional data. *Biometrika* **2015**, *102*, 421–437. https://doi.org/10.1093/biomet/asv006.
32. Shi, H.; Yang, Y.; Wang, L.; Ma, D.; Beg, M.F.; Pei, J.; Cao, J. Two-Dimensional Functional Principal Component Analysis for Image Feature Extraction. *Journal of Computational and Graphical Statistics* **2022**, *31*, 1127–1140. https://doi.org/10.1080/10618600.2022.2035738.
33. Van Der Linde, A. Variational Bayesian functional PCA. *Computational Statistics & Data Analysis* **2008**, *53*, 517–533. https://doi.org/10.1016/j.csda.2008.09.015.
34. Montagna, S.; Tokdar, S.T.; Neelon, B.; Dunson, D.B. Bayesian Latent Factor Regression for Functional and Longitudinal Data. *Biometrics* **2012**, *68*, 1064–1073. https://doi.org/10.1111/j.1541-0420.2012.01788.x.
35. Kowal, D.R.; Bourgeois, D.C. Bayesian Function-on-Scalars Regression for High-Dimensional Data. *Journal of Computational and Graphical Statistics* **2020-7-2**, *29*, 629–638. https://doi.org/10.1080/10618600.2019.1710837.
36. Sousa, P.H.T.O.; Souza, C.P.E.d.; Dias, R. Bayesian adaptive selection of basis functions for functional data representation. *Journal of Applied Statistics* **2024-04-03**, *51*, 958–992. https://doi.org/10.1080/02664763.2023.2172143.
37. Li, Y.; Wang, N.; Carroll, R.J. Selecting the Number of Principal Components in Functional Data. *Journal of the American Statistical Association* **2013**, *108*, 1284–1294. https://doi.org/10.1080/01621459.2013.788980.
38. Huang, L.; Reiss, P.T.; Xiao, L.; Zipunnikov, V.; Lindquist, M.A.; Crainiceanu, C.M. Two-way principal component analysis for matrix-variate data, with an application to functional magnetic resonance imaging data. *Biostatistics* **2017**, *18*, 214–229. https://doi.org/10.1093/biostatistics/kxw040.
39. Lynch, B.; Chen, K. A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika* **2018**, *105*, 815–831. https://doi.org/10.1093/biomet/asy048.
40. Shamshoian, J.; Şentürk, D.; Jeste, S.; Telesca, D. Bayesian analysis of longitudinal and multidimensional functional data. *Biostatistics* **2022/04/13**, *23*, 558–573. https://doi.org/10.1093/biostatistics/kxaa041.
41. Huo, S.; Morris, J.S.; Zhu, H. Ultra-Fast Approximate Inference Using Variational Functional Mixed Models. *Journal of Computational and Graphical Statistics* **2023-4-3**, *32*, 353–365. https://doi.org/10.1080/10618600.2022.2107532.
42. Jolliffe, I.T.; Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2016**, *374*, 20150202. https://doi.org/10.1098/rsta.2015.0202.
43. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **1999**, *61*, 611–622.
44. Ilin, A.; Raiko, T. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research* **2010**, *11*, 1957–2000.
45. Bishop, C. Bayesian pca. *Advances in neural information processing systems* **1998**, *11*.
46. Tipping, M.E.; Bishop, C.M. Mixtures of probabilistic principal component analyzers. *Neural computation* **1999**, *11*, 443–482. https://doi.org/10.1162/089976699300016728.
47. Minka, T. Automatic choice of dimensionality for PCA. *Advances in neural information processing systems* **2000**, *13*.

 doi:10.20944/preprints202503.0658.v1

48. Li, J.; Tao, D. On Preserving Original Variables in Bayesian PCA With Application to Image Analysis. *IEEE Transactions on Image Processing* **2012**, *21*, 4830–4843. https://doi.org/10.1109/TIP.2012.2211372.
49. Bouveyron, C.; Latouche, P.; Mattei, P.A. Bayesian variable selection for globally sparse probabilistic PCA. *Electronic Journal of Statistics* **2018**, *12*. https://doi.org/10.1214/18-EJS1450.
50. Ning, Y.C.B.; Ning, N. Spike and slab Bayesian sparse principal component analysis. *Statistics and Computing* **2024**, *34*, 118.
51. Bhattacharya, A.; Dunson, D.B. Sparse Bayesian infinite factor models. *Biometrika* **2011**, *98*, 291–306. https://doi.org/10.1093/biomet/asr013.
52. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *Journal of machine learning research* **2001**, *1*, 211–244.
53. Tipping, M. The relevance vector machine. *Advances in neural information processing systems* **1999**, *12*.
54. Faul, A.C.; Tipping, M.E., Analysis of Sparse Bayesian Learning. In *Advances in Neural Information Processing Systems 14*; The MIT Press, 2002; pp. 383–390.
55. MacKay, D.J., Bayesian methods for backpropagation networks. In *Models of neural networks III: association, generalization, and representation*; Springer, 1996; pp. 211–254.
56. Neal, R.M. *Bayesian learning for neural networks*; Vol. 118, Springer Science & Business Media, 2012.
57. Palmer, J.; Rao, B.; Wipf, D. Perspectives on sparse Bayesian learning. *Advances in neural information processing systems* **2003**, *16*.
58. Wipf, D.; Nagarajan, S. A new view of automatic relevance determination. *Advances in neural information processing systems* **2007**, *20*.
59. Wipf, D.; Rao, B. ℓ_0-norm minimization for basis selection. In Proceedings of the Proceedings of the 17th International Conference on Neural Information Processing Systems, 2004, pp. 1513–1520.
60. Wipf, D.P.; Rao, B.D.; Nagarajan, S. Latent Variable Bayesian Models for Promoting Sparsity. *IEEE Transactions on Information Theory* **2011**, *57*, 6236–6255. https://doi.org/10.1109/TIT.2011.2162174.
61. Wipf, D.; Rao, B. Sparse Bayesian Learning for Basis Selection. *IEEE Transactions on Signal Processing* **2004**, *52*, 2153–2164. https://doi.org/10.1109/TSP.2004.831016.
62. Bishop, C.M. Variational Principal Components. In Proceedings of the Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99. IEE, 1999, pp. 509–514.
63. Nakajima, S.; Sugiyama, M.; Babacan, D. On Bayesian PCA: Automatic dimensionality selection and analytic solution. In Proceedings of the Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 497–504.
64. Girolami, M.; Rogers, S. Hierarchic Bayesian models for kernel learning. In Proceedings of the Proceedings of the 22nd international conference on Machine learning, 2005, pp. 241–248.
65. Bishop, C.M.; Tipping, M.E. Variational relevance vector machines. In Proceedings of the Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, 2000, pp. 46–53.
66. Cheng, L.; Yin, F.; Theodoridis, S.; Chatzis, S.; Chang, T.H. Rethinking Bayesian Learning for Data Analysis: The Art of Prior and Inference in Sparsity-Aware Modeling. *IEEE Signal Processing Magazine* **2022**, *39*, 18–52. https://doi.org/10.1109/MSP.2022.3198201.
67. Qi, Y.A.; Minka, T.P.; Picard, R.W.; Ghahramani, Z. Predictive automatic relevance determination by expectation propagation. In Proceedings of the Twenty-first international conference, Banff, Alberta, Canada, 2004; p. 85. https://doi.org/10.1145/1015330.1015418.
68. Babacan, S.D.; Luessi, M.; Molina, R.; Katsaggelos, A.K. Sparse Bayesian Methods for Low-Rank Matrix Estimation. *IEEE Transactions on Signal Processing* **2012**, *60*, 3964–3977. https://doi.org/10.1109/TSP.2012.2197748.
69. Zhao, Q.; Meng, D.; Xu, Z.; Zuo, W.; Yan, Y. L_1 -Norm Low-Rank Matrix Factorization by Variational Bayesian Method. *IEEE Transactions on Neural Networks and Learning Systems* **2015**, *26*, 825–839. https://doi.org/10.1109/TNNLS.2014.2387376.
70. Guan, Y.; Dy, J. Sparse probabilistic principal component analysis. In Proceedings of the Artificial Intelligence and Statistics. PMLR, 2009, pp. 185–192.
71. Wong, A.P.S.; Wijffels, S.E.; Riser, S.C.; Pouliquen, S.; Hosoda, S.; Roemmich, D.; Gilson, J.; Johnson, G.C.; Martini, K.; Murphy, D.J.; et al. Argo Data 1999-2019: Two Million Temperature-Salinity Profiles and Subsurface Velocity Observations From a Global Array of Profiling Floats. *Frontiers in Marine Science* **2020**, *7*, 700. https://doi.org/ARTN70010.3389/fmars.2020.00700.

 doi:10.20944/preprints202503.0658.v1

72. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **2017**, *112*, 859–877. https://doi.org/10.1080/01621459.2017.1285773.
73. Tipping, M.E.; Faul, A.C. Fast marginal likelihood maximisation for sparse Bayesian models. In Proceedings of the International workshop on artificial intelligence and statistics. PMLR, 2003, pp. 276–283.
74. Park, T.; Lee, S. Improving the Gibbs sampler. *Wiley Interdisciplinary Reviews: Computational Statistics* **2022**, *14*, e1546.
75. Ledoit, O.; Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **2004**, *88*, 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4.
76. Kaslow, R.A.; Ostrow, D.G.; Detels, R.; Phair, J.P.; Polk, B.F.; RINALDO Jr, C.R.; Study, M.A.C. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *American journal of epidemiology* **1987**, *126*, 310–318. https://doi.org/10.1093/aje/126.2.310.
77. Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC of November 09st 2024, 2024. https://doi.org/10.17882/42182.
78. Yarger, D.; Stoev, S.; Hsing, T. A functional-data approach to the Argo data. *The Annals of Applied Statistics* **2022**, *16*, 216–246. https://doi.org/10.1214/21-Aoas1477.
79. de Boyer Montégut, C.; Madec, G.; Fischer, A.S.; Lazar, A.; Iudicone, D. Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. *Journal of Geophysical Research: Oceans* **2004**, *109*.
80. Roemmich, D.; Gilson, J. The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. *Progress in oceanography* **2009**, *82*, 81–100. https://doi.org/10.1016/j.pocean.2009.03.004.
81. Kuusela, M.; Stein, M.L. Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proceedings of the Royal Society A* **2018**, *474*, 20180400. https://doi.org/10.1098/rspa.2018.0400.