Efficient Multidimensional Functional Data Analysis Using Marginal Product Basis Systems

William Consagra, Arun Venkataraman, and Xing Qiu

Abstract

In areas ranging from neuroimaging to climate science, advances in data storage and sensor technology have led to a proliferation in multidimensional functional datasets. A common approach to analyzing functional data is to first map the discretely observed functional samples into continuous representations, and then perform downstream statistical analysis on these smooth representations. It is well known that many of the traditional approaches used for 1D functional data representation are plagued by the curse of dimensionality and quickly become intractable as the dimension of the domain increases. In this article, we propose a computational framework for learning continuous representations from a sample of multidimensional functional data that is immune to several manifestations of the curse. The representations are constructed using a set of separable basis functions that are defined to be optimally adapted to the data. We show that the resulting estimation problem can be solved efficiently by the tensor decomposition of a carefully defined reduction transformation of the observed data. Roughness-based regularization is incorporated using a class of differential operator-based penalties. Relevant theoretical properties are also discussed. The advantages of our method over competing methods are thoroughly demonstrated in simulations. We conclude with a real data application of our method to a clinical diffusion MRI dataset.

1. Introduction

Functional data analysis (FDA) is a subfield of statistics concerned with the analysis of samples of functions. In most applications, the functional sample is not observed directly, rather, at some discrete number of domain points \( x_{ij} \in \mathcal{M} \subset \mathbb{R}^D \) according to

\[
Y_{ij} = U_i(x_{ij}) + \epsilon_{ij}; \quad i = 1, \ldots, N; j = 1, \ldots, M_i
\]

where \( N \) is the sample size, \( M_i \) the number of domain points for the \( i \)-th sample, \( U_i \sim U \) is a random function and \( \epsilon_{ij} \) is additive noise. In many FDA workflows, the analyst needs to perform the initial step of estimating a smooth function \( \hat{U}_i \) from each subject’s discretely observed noisy data, that is, the “smoothing first, then estimation” approach (Zhang and Chen 2007), which we refer to here as functional representation. Various downstream analyses are then performed using these reconstructed smooth functions. In this work, we are interested in the function representation problem for multidimensional functional data, that is, random functions defined on multidimensional (\( D \geq 1 \)) domains. Specifically, we consider the situation where the sample is observed on a common \( D \)-dimensional grid, that is, \( x_{ij} \) are the same for all subjects, a setting frequently encountered in various imaging applications.

When \( D = 1 \), functional representation can be accomplished using standard nonparametric approaches, for example, local kernel regression or basis expansion (Ramsay and Silverman 2005; Hsing and Eubank 2015). For \( D > 1 \), these approaches suffer from the curse of dimensionality: the number of observations required to obtain a desired mean-squared error (Stone 1980) and/or the number of model parameters (e.g., the number of basis functions for a tensor product basis) grow exponentially in \( D \) (Wasserman 2010). In such a scenario, semiparametric regression approaches provide a route to tractable estimation, though the associated structural assumptions are often overly restrictive for real word data. Hence, developing a general framework for multidimensional functional data analysis demands a different approach to representation.

It is well known that the optimal (minimum mean integrated squared error) low-rank representation for functional data can be formed using the eigenfunctions of the covariance operator of the process, and thus these form an attractive basis system for parsimonious modeling. Unfortunately, estimating the eigenfunctions, that is, performing functional principal component analysis (FPCA) (Silverman 1996), for a general \( D \)-dimensional random function requires the nonparametric estimation of the \( 2D \)-dimensional covariance function, denoted \( C(x,y) \). Some general techniques have recently been proposed (Chen and Jiang 2017; Li, Huang, and Härdle 2019; Wang, Wong, and Zhang 2020), though the curse of dimensionality is even more problematic in this situation, due to both the dimension doubling effect and the symmetry and positive semidefinite constraints. To
alleviate the computational difficulties associated with estimating a generic covariance function of a multidimensional process, a common tactic is to assume some notion of separability for $C$ (Chen, Delicado, and Müller 2017; Lynch and Chen 2018). These works are mostly developed for the case when the domain naturally decomposes into a product of two spaces, for example, space and time, though extending the theory to general $D$-dimensional domains is feasible. Computationally speaking, it is a different story. For instance, when $D > 2$, the marginal product FPCA in Chen, Delicado, and Müller (2017) requires multidimensional numerical integration or nonparametric smoothing in order to estimate the marginal covariance functions, reintroducing a manifestation of the curse of dimensionality.

In what follows, we propose a framework for multidimensional function representation based on learning the optimal marginal product basis (MPB), that is, a collection of independent multiplicatively separable functions, that avoids both direct estimation of, and explicit structural assumptions on, $C$. Critically, the number of parameters needed to estimate the MPB is linear in $D$, and therefore this structure is effective for combating the curse of dimensionality. We prove that the optimal MPB defines a representation space that can be considered nearly optimal for a particular rank, with an inefficiency cost that becomes negligible for large ranks. To estimate the optimal MPB from the observed data, we identify an isometric embedding which allows the reparameterization of the observed data tensor into a lower dimensional space, permitting the derivation of a fast algorithm which scales favorably with huge datasets. This is in contrast to alternative methods for optimal basis construction for multidimensional functional data which rely on a smoothed decomposition of the raw data tensor (Huang, Shen, and Buja 2009; Allen 2013; Allen and Weylandt 2019), and as a result become computationally problematic for densely observed functions. Additionally, our approach enables a “fully functional” treatment of the estimation problem, in which the continuous basis functions are estimated directly, as opposed to the discretization of roughness-based regularization using differential operators.

2. Marginal Product Models

2.1. Background and Model Description

Let $U$ be a multidimensional random function with real-valued square integrable realizations, that is, $U_i \sim U \in \mathcal{H} := L^2(\mathcal{M})$. We assume the domain can be decomposed as $\mathcal{M} = M_1 \times \cdots \times M_D$, so for $x = (x_1, \ldots, x_D) \in \mathcal{M}$, each $x_d$ for $d = 1, \ldots, D$, is a member in the marginal domain $M_d$, which is assumed to be a compact subset of Euclidean space $\mathbb{R}^d$. Without loss of generality, we assume the Lebesgue measure of $M$ is 1. We make the following regularity assumptions on $U$:

\textbf{Assumption 1.} (a) $\mathbb{E}[U] = 0$, (b) $\mathbb{E} \int_{\mathcal{M}} U^2(x) \, dx < \infty$, and (c) $U$ is mean-square continuous.

The mean zero assumption on $U$ is made for convenience of presentation, and the mean square integrability and mean-square continuity assumptions are a standard requirement (Hsing and Eubank 2015). Under Assumption 1, we are guaranteed that the covariance function $C(x, y) := \mathbb{E}[U(x)U(y)]$ is continuous on $\mathcal{M} \times \mathcal{M}$. By Mercer’s theorem, this covariance function has an eigen-decomposition

\[ C(x, y) = \sum_{k=1}^{\infty} \rho_k \psi_k(x) \psi_k(y), \]

where $\{\psi_k\}_{k=1}^{\infty}$ forms a complete orthonormal sequence of eigenfunctions in $\mathcal{H}$ and $\{\rho_k\}_{k=1}^{\infty}$ is a nonincreasing sequence of real, nonnegative eigenvalues. Additionally, by the Karhunen-Loève theorem, with probability one we have the decomposition $U(x) = \sum_{k=1}^{\infty} Z_k \psi_k(x)$, where $Z_k = \langle U, \psi_k \rangle_{\mathcal{H}}$, which are mean zero random variables with $\mathbb{E}[Z_k Z_j] = \delta_{k,j} \rho_k$. Let $H_d := L^2(M_d)$, so that $\mathcal{H} := L^2(\mathcal{M}) = \bigotimes_{d=1}^{D} H_d$, the tensor product of $D$ member spaces. We assume that there exists a complete basis system, $\phi_d := \{\phi_{d,j}\}_{j=1}^{\infty}$, for each marginal function space $H_d$. Denote their rank-$m_d$ truncations as $\phi_{m_d,d} := \{\phi_{d,1}, \ldots, \phi_{d,m_d}\}$; $H_{m_d,d} := \text{span}(\phi_{m_d,d})$; and $\mathcal{H}_m := \bigotimes_{d=1}^{D} H_{m_d,d}$, where $m = (m_1, \ldots, m_D)$ are the marginal ranks. By construction,

\[ \tau_m := \bigotimes_{d=1}^{D} \phi_{m_d,d} = \left\{ \tau_{j_1, \ldots, j_D}(x) = \prod_{d=1}^{D} \phi_{d,j_d}(x_d), j_d = 1, \ldots, m_d \right\} \]

is the complete tensor product bases (TPB) for $\mathcal{H}_m$.  

The rest of the article is organized as follows. In Section 2, we formulate the optimal MPB system and discuss relevant theoretical properties. In Section 3, we derive an efficient estimation procedure and discuss the incorporation of roughness-based regularization using differential operators as well as hyperparameter selection. In Section 4, we illustrate how to use the MPB to derive a fast two-stage multidimensional FPCA. Section 5 compares the proposed method with competing methods in simulation studies. In Section 6, we analyze a set of magnetic resonance imaging data from a traumatic brain injury study. Section 7 offers concluding remarks and potential future directions. Proofs for all theorems can be found in supplemental materials. Code for our algorithms and scripts to reproduce all of the results in the simulation section have been made publicly available in a python package: https://github.com/Will-Consagra/eMFDA.
Definition 2.1 (Marginal product structure). \( \xi \in \mathcal{H} \) is called a rank-1 marginal product function if it is multiplicative separable, and \( u \in \mathcal{H} \) is called a rank-\( K \) marginal product function if it is a linear combination of \( K \) independent rank-1 marginal product functions:

\[
u(x) = \sum_{k=1}^{K} b_k \xi_k(x) = \sum_{k=1}^{K} b_k \prod_{d=1}^{D} \xi_{k,d}(x_d), \quad \xi_{k,d} \in H_d, \quad b_k \in \mathbb{R}.
\]

We denote the collection of rank-1 marginal product functions with marginal ranks \( m \):

\[
\mathcal{L}_m := \{ \xi(x) : \xi(x) = \prod_{d=1}^{D} \xi_d(x_d), \xi_d \in H_{m_d,d}, \|\xi_d\|_{H_d} = 1 \}.
\]

In this work, we propose to estimate the optimal basis set of \( K \) elements from \( \mathcal{L}_m \) for representing realizations of \( U \), a notion formalized as follows:

Definition 2.2 (Optimal Rank-\( K \) MPB). Define the set of functions

\[
\mathcal{V}_{K,m} := \{ \xi = (\xi_1, \ldots, \xi_K) : \xi_k \in \mathcal{L}_m, \xi_1, \ldots, \xi_K \text{ linearly independent} \},
\]

and define the associated optimal rank \( K \) MPB, denoted \( K \)-oMPB, as

\[
\xi^*_m = \arg \inf_{\xi \in \mathcal{V}_{K,m}} \mathbb{E} \left\| U - P_\xi(U) \right\|_{\mathcal{H}}^2,
\]

where \( P_\xi \) is the projection operator onto span(\( \xi \)).

Given a random sample of \( N \) realizations \( U_i \sim U \), define the corresponding empirical estimate of (4) as

\[
\tilde{\xi}^*_m,N = \arg \inf_{\xi \in \mathcal{V}_{K,m}} \frac{1}{N} \sum_{i=1}^{N} \left\| U_i - P_\xi(U_i) \right\|_{\mathcal{H}}^2.
\]

2.2. Approximation Properties

We now characterize the expected asymptotic approximation power of \( \tilde{\xi}^*_m,N \). Let \( A_k \) be the \( D \)-mode tensor with elements \( A_k(j_1, \ldots, j_D) \) defined by

\[
P_{H_m}(\psi_k) = \sum_{j_1=1}^{m_1} \cdots \sum_{j_D=1}^{m_D} A_k(j_1, \ldots, j_D) \phi_{1,j_1} \cdots \phi_{D,j_D},
\]

where \( P_{H_m} \) is the projection operator onto \( H_m \). Denote \( A^{(K)} \) as the tensor obtained by the \( D + 1 \) mode stacking of \( A_1, \ldots, A_K \). Let \( J_{\phi_d} \in \mathbb{R}^{m_d \times m_d} \) be the matrix of pairwise \( H_{m_d,d} \)-inner products of \( \phi_{m_d,d} \). Define the inner-product space \( (\otimes_{d=1}^{D} \mathbb{R}^{m_d} \otimes \mathbb{R}^{K}, \langle \cdot, \cdot \rangle_{F,C}) \), where

\[
\langle T_1, T_2 \rangle_{F,C} = \sum_{k=1}^{K} \rho_k (T_1(\ldots, \cdot, k), T_2(\ldots, \cdot, k) \times J_{\phi_d} )
\]

for tensors \( T_1, T_2 \in \otimes_{d=1}^{D} \mathbb{R}^{m_d} \otimes \mathbb{R}^{K} \), where \( x_d \) denotes the tensor \( d \)-mode multiplication.

Theorem 2.1 (Generalization Error). With slight abuse of notation, denote \( P_{\tilde{\xi}^*_{m,N}} \) as the projection operator onto span(\( \tilde{\xi}^*_{m,N} \)) and let the function \( w_{\tau_m}(m) \) be the \( L^2(M) \) convergence rate of the TPB system \( \tau_m \) (Definition S1.1 in the supplementary materials). Under Assumptions 1 and S2, S3 in the supplementary material,

\[
\mathbb{E} \left\| U - P_{\tilde{\xi}^*_{m,N}}(U) \right\|_{\mathcal{H}}^2 \leq \sum_{k=K+1}^{\infty} \rho_k + \left\| \mathcal{A}^{(K)} - \mathcal{A}^{(K)}_{\text{op}} \right\|_{F,C}^2 + O(w_{\tau_m}(m)) + O_p(N^{-1/2})
\]

where the expectation is taken with respect to a new realization of \( U \) and \( \mathcal{A}^{(K)}_{\text{op}} \) is the rank \( K \) canonical polyadic decomposition of the coefficient tensor \( \mathcal{A}^{(K)} \) under the \( \| \cdot \|_{F,C} \)-norm.

Theorem 2.1 bounds the expected generalization error of \( \tilde{\xi}^*_{m,N} \) by the sum of four terms. The first term is the tail sum of the eigenvalues, which is the expected generalization error of the optimal rank \( K \) basis system (the eigenfunctions). The third term is the irreducible bias from the finite truncation of the marginal ranks \( m \). The fourth term reflects the finite sample statistical approximation error and can be established using convergence results from the theory of M-estimators. The second term shows that the inefficiency cost incurred by representing the function realizations with \( \tilde{\xi}^*_{m,N} \), as opposed to the eigenfunctions, is driven by the low-rank structure of \( \mathcal{A}^{(K)} \). This term is unknown in practice but vanishes for large enough \( K \) (often, \( K \ll \prod_{d=1}^{D} m_d \)) though this value will depend on the particular tensor product space \( \tau \), marginal ranks \( m \) and covariance function \( C \). Please visit Section S1 of the supplementary materials for further technical discussion on these and related theoretical matters. We now turn our attention to the development of computationally efficient algorithms to estimate \( \tilde{\xi}^*_{m,N} \) in practice.

3. Estimation

3.1. Discrete Observation Model

We consider the case in which the \( U_i \) are observed with noise at each discrete location of a multidimensional grid \( \mathcal{X} \subset \mathcal{M} \), where

\[
\mathcal{X} = (x_{11}, x_{12}, \ldots, x_{1n_1}) \times (x_{21}, x_{22}, \ldots, x_{2n_2}) \times \cdots \times (x_{D1}, x_{D2}, \ldots, x_{Dn_D}),
\]

and each vector of marginal grid points \( x_d := (x_{d1}, x_{d2}, \ldots, x_{dn_d}) \in \mathcal{M}_d \), according to the canonical observation model

\[
\mathcal{Y}(i_1, i_2, \ldots, i_D, i) = U_j(x_{1,i_1}, x_{2,i_2}, \ldots, x_{D,i_D}) + \epsilon(i_1, i_2, \ldots, i_D, i)
\]

for \( i_d = 1, 2, \ldots, n_d, d = 1, 2, \ldots, D, i = 1, 2, \ldots, N, \) where \( \mathcal{Y} \) is a \( D + 1 \)-mode tensor with dimensions \( (n_1, n_2, \ldots, n_D, N) \), \( \mathbb{E}[\vec{\epsilon}(\mathcal{X})] = 0 \) and \( \text{var}(|\epsilon(\mathcal{X})|) = \sigma^2 I \). The discretized counterpart to (5) is given by

\[
\tilde{\xi}^*_{N,m} := \arg \inf_{\xi \in \mathcal{V}_{K,m}} \min_{B_{k,l}} \frac{1}{N} \sum_{i=1}^{N} \left\| \mathcal{Y}_i - \sum_{k=1}^{K} \sum_{d=1}^{D} B_{k,l} \otimes \tilde{\xi}_{d,k} \right\|_{F}^2,
\]
where \( Y_i \in \mathbb{R}^{n_1 \times \cdots \times n_D} \) is the observed data tensor for the \( i \)th realization, \( \xi_{d,k} \in \mathbb{R}^{n_d} \) is the evaluation of \( \xi_{k,d} \) on \( x_d \) and \( B \) is the matrix of coefficients for the \( \xi_k \)'s for each of the \( N \) samples.

### 3.2. A Convenient Reparameterization

First we note that for \( \xi \in \mathcal{V}_{K,m} \), we have the representation \( \xi_{k,d}(x_d) = \sum_{j=1}^{m_d} c_{d,k,j} \phi_{d,k}(x_d) \). Consequently, \( \xi_{N,m} \) is equivalently defined by the solutions to the following optimization problem

\[
(\hat{C}_1, \ldots, \hat{C}_D) := \arg \min_{(C_1, \ldots, C_D)} \frac{1}{N} \sum_{i=1}^{N} \left\| y_i - \sum_{k=1}^{K} \sum_{d=1}^{D} B_{ik} \otimes c_{d,k} \right\|_F^2
\]

for some \( \lambda_d > 0 \), where \( L_d : \mathbb{R}^{n_d \times \alpha(\mathcal{M}_d)} \to \mathbb{R}^{2(\mathcal{M}_d)} \) is a linear (partial) differential operator, and \( \mathcal{V}_{m_d \times \alpha(\mathcal{M}_d)} \) is the Sobolev space over the \( d \)th marginal domain, with order \( \alpha_d \) defined appropriately.

**Proposition 3.2.** Let \( T_d := D_d^{-1} V_d' R_d V_d D_d^{-1} \), with \( R_d(i,j) = \int_{\mathcal{M}_d} L_d(\phi_{d,i}) L_d(\phi_{d,j}) \), then \( \sum_{k=1}^{K} \operatorname{Pen}(\xi_k) = \sum_{d=1}^{D} \lambda_d \operatorname{tr}(C_d' T_d C_d) \).

As a consequence, such penalties are quadratic in the transformed coordinate matrices \( \hat{C}_d \) and therefore convex. This permits the derivation of efficient numerical algorithms to estimate the optimal MPB functions, which will be discussed in Section 3.4.

Penalization of the coefficient matrix \( B \) is incorporated using the penalty function denoted \( l(B) \). We assume that \( l(B) \) is convex, which is a requirement to guarantee the convergence of the algorithm in Section 3.4, but otherwise leave it’s form unspecified. For example, lasso-type penalties can be integrated to promote sparsity in the MPB functional representation for interpretability or feature extraction. Using the results from Theorem 3.1 and Proposition 3.2, the solution to the regularized augmentation of (9) is a linear transformation of the solution to

\[
\arg \min_{\hat{C}_1, \ldots, \hat{C}_D, B} \left\| \tilde{G} - \sum_{k=1}^{K} \sum_{d=1}^{D} \tilde{c}_{d,k} \otimes b_k \right\|_F^2 + \sum_{d=1}^{D} \lambda_d \operatorname{tr}(\tilde{C}_d' T_d \tilde{C}_d) + \lambda_{D+1} l(B).
\]

### 3.4. Algorithm

In general, it can be shown that the optimization problem (10) is nonconvex and NP-hard (Hillar and Lim 2013). To derive a computationally tractable approximation algorithm, we propose a block coordinate descent based approach in which, for the \((r + 1)\)th iteration, the variables are updated according to the sequence of conditional minimization problems

\[
\hat{C}_d^{(r+1)} = \min_{\chi} \hat{G}_1^{(r+1)} \ldots, \hat{C}_{d-1}^{(r+1)}, X, \hat{C}_{d+1}^{(r)}, \ldots, \hat{C}_D^{(r), B^{(r)}},
\]

for \( d = 1, \ldots, D \) and likewise for \( B^{(r+1)} \), where \( g \) denotes the objective function from (10).

Using the properties of the \( d \)-mode matricization, we can write the conditional minimization problem defining the update of \( \hat{C}_d \) as

\[
\hat{C}_d^{(r+1)} = \min_{\hat{C}_d} \left\| G_{(d)}^{(r)} - \hat{C}_d W_d^{(r)} \right\|_F^2 + \lambda_d \operatorname{tr}(\tilde{C}_d' T_d \tilde{C}_d).
\]

The update for \( B \) is given by

\[
B^{(r+1)} = \min_B \left\| G_{(D+1)}^{(r)} - W_{D+1}^{(r)} B \right\|_F^2 + \lambda_{D+1} l(B),
\]

in order to ameliorate the influence of noise and discretization and to promote smoothness in the estimated K-OMPB basis, we incorporate a regularization term to the objective function in (9) of the form:

\[
\sum_{k=1}^{K} \operatorname{Pen}(\xi_k) = \sum_{k=1}^{K} \frac{1}{\lambda_d} \int_{\mathcal{M}_d} L_d(\phi_{d,k}) L_d(\phi_{d,k}) \]

for some \( \lambda_d > 0 \), where \( L_d : \mathbb{R}^{n_d \times \alpha(\mathcal{M}_d)} \to \mathbb{R}^{2(\mathcal{M}_d)} \) is a linear (partial) differential operator, and \( \mathcal{V}_{m_d \times \alpha(\mathcal{M}_d)} \) is the Sobolev space over the \( d \)th marginal domain, with order \( \alpha_d \) defined appropriately.

### 3.3. Regularization in the Transformed Space

In order to ameliorate the influence of noise and discretization and to promote smoothness in the estimated K-OMPB basis, we incorporate a regularization term to the objective function in (9) of the form:

\[
\sum_{k=1}^{K} \operatorname{Pen}(\xi_k) = \sum_{k=1}^{K} \frac{1}{\lambda_d} \int_{\mathcal{M}_d} L_d(\phi_{d,k}) L_d(\phi_{d,k}) \]

for some \( \lambda_d > 0 \), where \( L_d : \mathbb{R}^{n_d \times \alpha(\mathcal{M}_d)} \to \mathbb{R}^{2(\mathcal{M}_d)} \) is a linear (partial) differential operator, and \( \mathcal{V}_{m_d \times \alpha(\mathcal{M}_d)} \) is the Sobolev space over the \( d \)th marginal domain, with order \( \alpha_d \) defined appropriately.

**Proposition 3.2.** Let \( T_d := D_d^{-1} V_d' R_d V_d D_d^{-1} \), with \( R_d(i,j) = \int_{\mathcal{M}_d} L_d(\phi_{d,i}) L_d(\phi_{d,j}) \), then \( \sum_{k=1}^{K} \operatorname{Pen}(\xi_k) = \sum_{d=1}^{D} \lambda_d \operatorname{tr}(C_d' T_d C_d) \).

As a consequence, such penalties are quadratic in the transformed coordinate matrices \( \hat{C}_d \) and therefore convex. This permits the derivation of efficient numerical algorithms to estimate the optimal MPB functions, which will be discussed in Section 3.4.

Penalization of the coefficient matrix \( B \) is incorporated using the penalty function denoted \( l(B) \). We assume that \( l(B) \) is convex, which is a requirement to guarantee the convergence of the algorithm in Section 3.4, but otherwise leave it’s form unspecified. For example, lasso-type penalties can be integrated to promote sparsity in the MPB functional representation for interpretability or feature extraction. Using the results from Theorem 3.1 and Proposition 3.2, the solution to the regularized augmentation of (9) is a linear transformation of the solution to

\[
\arg \min_{\hat{C}_1, \ldots, \hat{C}_D, B} \left\| \tilde{G} - \sum_{k=1}^{K} \sum_{d=1}^{D} \tilde{c}_{d,k} \otimes b_k \right\|_F^2 + \sum_{d=1}^{D} \lambda_d \operatorname{tr}(\tilde{C}_d' T_d \tilde{C}_d) + \lambda_{D+1} l(B).
\]

### 3.4. Algorithm

In general, it can be shown that the optimization problem (10) is nonconvex and NP-hard (Hillar and Lim 2013). To derive a computationally tractable approximation algorithm, we propose a block coordinate descent based approach in which, for the \((r + 1)\)th iteration, the variables are updated according to the sequence of conditional minimization problems

\[
\hat{C}_d^{(r+1)} = \min_{\chi} \hat{G}_1^{(r+1)} \ldots, \hat{C}_{d-1}^{(r+1)}, X, \hat{C}_{d+1}^{(r)}, \ldots, \hat{C}_D^{(r), B^{(r)}},
\]

for \( d = 1, \ldots, D \) and likewise for \( B^{(r+1)} \), where \( g \) denotes the objective function from (10).

Using the properties of the \( d \)-mode matricization, we can write the conditional minimization problem defining the update of \( \hat{C}_d \) as

\[
\hat{C}_d^{(r+1)} = \min_{\hat{C}_d} \left\| G_{(d)}^{(r)} - \hat{C}_d W_d^{(r)} \right\|_F^2 + \lambda_d \operatorname{tr}(\tilde{C}_d' T_d \tilde{C}_d).
\]

The update for \( B \) is given by

\[
B^{(r+1)} = \min_B \left\| G_{(D+1)}^{(r)} - W_{D+1}^{(r)} B \right\|_F^2 + \lambda_{D+1} l(B),
\]
where $W_d^{(r)} = (\bigodot_{j=1}^D \tilde{C}_d^{(r+1)} \bigodot_{j=1}^D \tilde{C}_d^{(r)}) \odot B^{(r)}$ for $d = 1, \ldots, D$, $W_{d+1}^{(r)} = \bigodot_{d=1}^D \tilde{Z}_d^{(r+1)} + G(d)$ is the $d$-mode unfolding of tensor $\tilde{G}$. Here $\bigodot$ is the Khatri-Rao product. From here on the superscript $r$ denoting iteration is dropped for clarity.

In fact, the solution to the subproblem (12) is equivalent to the solution to

$$\tilde{C}_d W_d^{(r)} W_d + \lambda_d T_d \tilde{C}_d = W_d^{(r)} G(d),$$

This equivalence can be verified by noting that (14) defines the gradient equations of (12) and that the solution is globally optimum due to convexity. Equation (14) is known as the Sylvester equation and has a unique solution under very mild conditions (specifically $W_d^{(r)}$ and $\lambda_d T_d$ must have no common eigenvalues). Efficient algorithms for solving the Sylvester equation (Bartels and Stewart 1972) are readily available in most common numerical computing languages.

Notice that by introducing the auxiliary variable $Z = B^{*}$, the subproblem (13) can be written in separable form as

$$\min_{BZ} \|G(D+1) - W_{D+1} Z\|_F^2 + \lambda_{D+1} l(B) \quad \text{subject to } B - Z = 0, \quad (15)$$

A numerical approximation to problems of the form (15) can be found using an alternating direction method of multipliers (ADMM) algorithm (Boyd et al. 2011). The ADMM scheme consists of the iterates

$$B_{\text{update}} \leftarrow \min_B \left( \lambda_{D+1} l(B) + \gamma \|B - Z^* + A^*\|_F^2 \right),$$

$$Z_{\text{update}} \leftarrow \min_Z \left( \|G(D+1) - W_{D+1} Z\|_F^2 + \gamma \|B - Z^* + A^*\|_F^2 \right),$$

$$A^*_{\text{update}} \leftarrow A^* + B - Z^*$$

for some choice of $\gamma > 0$, where $A^*$ is the scaled dual variable associated with the constraint. Since $l$ is assumed to be convex, the ADMM iterates are guaranteed to converge.

The update (17) is a matrix ridge regression and has analytic solution given by

$$Z_{\text{update}} = [W_{D+1}^{*} W_{D+1} + \gamma I]^{-1} [W_{D+1}^{*} G(D+1) + \gamma (B + A^*)^*].$$

The update (16) defines the so-called proximal operator of $l$ and is uniquely minimized. The exact solution will depend on the form of $l$, but it can be shown that many reasonable choices permit an analytic result. For example, if $l(\cdot) = \| \cdot \|_1$, the update is given by the element-wise soft thresholding operator applied to matrix $Z^* - A^*$.

Algorithm 1 in the Supplemental Material provides pseudocode for estimating the K-oMPB using the block coordinate descent scheme, referred to from here on as MARGARITA (MARGinal-product bAsis Representation with Tensor Analysis). The convergence of MARGARITA to a stationary point can be guaranteed if each of the sub-problems is convex and has a unique solution (Bertsekas 1997). The former property is satisfied by our construction, while the latter is difficult to verify in practice but can be enforced with minor augmentations. In particular, adding an additional proximal regularization of the form $\frac{\mu}{2} \|X - \widetilde{C}_d^{(r)}\|_F^2$, for $\mu > 0$ to (11) guarantees strong convexity, and hence convergence.

We conclude this section with several remarks on practical implementation. Forming the matrix products $W_d^{(r)} W_d$ and $W_d^{(r)} G(d)$ can become computationally expensive when $D$ and/or $m_d$ become sufficiently large. To avoid this computational bottleneck, the former can be calculated efficiently by leveraging the identity $[\bigodot, A_i]'[\bigodot, A_i] = \bigodot, A_i'A_i$, where $\bigodot$ is the Hadamard product. Algorithms for efficient computation of the latter have been developed, see Phan, Tichavský, and Cichocki (2013). Following the suggestion of Huang, Sidiropoulos, and Liavas (2016), we found success setting $\gamma = \|W_d^{(r)} W_{D+1}\|_F / K$. Finally, if $l(\cdot) = \| \cdot \|_2^2$, then it is easy to show that (13) has a closed form solution and thus the ADMM scheme need not be invoked for this special case.

### 3.5. Hyperparameter Selection

A distinct advantage of our methodology is its flexibility in allowing the user to incorporate different notions of smoothness, via linear differential operator $L_d$, different choices of marginal basis systems and alternative coefficient penalty methods. In some cases, these can be selected using a-priori knowledge of the problem of interest, though for many applications a data-driven approach to hyperparameter selection may be of interest or of necessity. While our method provides the flexibility of setting $\lambda_d$ independently for all $d = 1, \ldots, D + 1$, using a data-driven method to select all these parameters is computationally infeasible for even moderately large $D$. Therefore, absent a-priori knowledge of different behavior in different dimensions, we suggest setting $\lambda_d = \lambda_f$ for $d = 1, \ldots, D$, and selecting the parameters $(\lambda_f, \lambda_{D+1})$ by minimizing the $n$-fold cross-validation error over a 2-dimensional grid. Pseudocode for this scheme is provided in Algorithm 2 in the supplemental material. Our method also requires the specification of both marginal ranks $m$ and a global rank $K$. We propose the following two proportion of variance explained measures

$$\text{PVG}(K) = \|\mathbb{G} - \sum_{k=1}^K b_k \otimes (\bigodot_{d=1}^D \tilde{C}_{d,k})\|_F^2 / \|\mathbb{G}\|_F^2,$$

$$\text{PVM}(m) := \|\mathbb{Y} \times_{d=1}^D \mathbb{U}_d^{(r)}\|_F^2 / \|\mathbb{Y}\|_F^2,$$

which can be used along with an elbow type criteria for selection. A detailed elaboration, justification and numerical evaluation for the proposed hyperparameter selection criteria is provided in Section S3 of the supplemental text.

### 4. Multidimensional Penalized FPCA

In this section, we demonstrate how to leverage the MPB structure to define a fast multidimensional FPCA which avoids the curse of dimensionality, incurring only trivial additional computational expense beyond MARGARITA. Consider the method for FPCA proposed in Silverman (1996), in which the $j$th eigenfunction $\psi_j$ is defined as the function maximizing the penalized
sample variance with modified orthogonality constraints

\[ \hat{\psi}_j = \max_{\psi \in \mathbb{W}^{\alpha,2}(M)} \frac{\sum_{i=1}^N \text{var}(\langle \psi, U_i \rangle_{\mathcal{H}})}{\langle \psi, \psi \rangle_\lambda} \]

s.t. \( \|\psi\|_{\mathcal{H}}^2 = 1, \langle \psi, \psi \rangle_\lambda = 0, \) for \( k = 1, 2, \ldots, j - 1. \) \( \text{(20)} \)

Here \( \langle \psi, \psi \rangle_\lambda := \langle \psi, \psi \rangle_{\mathcal{H}} + \lambda(L(\psi_j), L(\psi_j))_{\mathcal{H}} \) and \( L : \mathbb{W}^{\alpha,2}(M) \rightarrow \mathcal{H} \) is an \( \alpha \)th order linear differential operator quantifying the global roughness. For simplicity, hereafter we define \( L := \Delta_M, \) the Laplacian operator on \( M, \) though other linear differential operators can be incorporated effortlessly. In our set-up, \( L \) facilitates the optional incorporation of a flexible global notion of smoothness in addition to the marginally independent regularization in \( \text{(10)} \), for example penalizing mixed partial derivatives.

In the one-dimensional case, the optimization problem \( \text{(20)} \) is solved using a two-stage approach: first computing \( \hat{U}_i \) through expansion over some suitable basis system and then looking for solutions \( \hat{\psi}_j \) in the span of that set of basis functions. Analogously, we can first represent the realizations with the \( K \)-oMPB: \( \hat{U}_i(x) = b_i^* \zeta_m^{(i)}(x), \) and then solve \( \text{(20)} \) with the additional constraint \( \psi_j(x) = \hat{s}_j \zeta_m^{(i)}(x) \) for some \( s_j \in \mathbb{R}^K. \) Under this setup, the optimization problem \( \text{(20)} \) is equivalent to

\[ s_j = \max_s \frac{s' J_{\mathcal{M}}^*_m \Sigma_b J_{\mathcal{M}}^* s}{s' J_{\mathcal{M}}^* s + \lambda s' R_{\mathcal{M}}^* s} \]

s.t. \( s' J_{\mathcal{M}}^* s = 1, \) \( s' J_{\mathcal{M}}^* + \lambda R_{\mathcal{M}}^* |s_k| = 0, \) for \( k = 1, 2, \ldots, j - 1. \) \( \text{(21)} \)

Here \( \Sigma_b = \text{cov}(b) \) and \( J_{\mathcal{M}}^* \) are symmetric PSD matrices with elements \( \langle J_{\mathcal{M}}^* \rangle_{i,j} = \langle \zeta_m^{(i)}, \zeta_m^{(j)} \rangle_{\mathcal{H}} \) and \( R_{\mathcal{M}}^* \) is a diagonal matrix with elements \( \langle \Delta_M (\zeta_m^{(i)}), \Delta_M (\zeta_m^{(j)}), \Delta_M (\zeta_m^{(j)}) \rangle_{\mathcal{H}}. \) The objective function in \( \text{(21)} \) is a generalized Rayleigh quotient and it can be shown that the solutions for \( j = 1, \ldots, K \) are equivalently defined by the first \( K \) solutions to the generalized eigenvalue problem

\[ J_{\mathcal{M}}^* \Sigma_b J_{\mathcal{M}}^* s_j = v_j J_{\mathcal{M}}^* s_j + \lambda R_{\mathcal{M}}^* |s_j|, \]

hence, the vector of estimated eigenfunctions is \( \hat{\psi}(x) := (s_1 \zeta_m^{(i)}(x), \ldots, s_j \zeta_m^{(i)}(x))'. \)

There are several algorithms to solve the generalized eigenvalue problem \( \text{(22)} \). In practice, we use Algorithm 9.4.2 in Ramsay and Silverman (2005), which requires the computation of the marginal inner product matrices: \( J_{\phi_d}(i,j) = \langle \phi_d(i), \phi_d(j) \rangle_{\mathcal{H}} \), \( R_{\phi_d}(i,j) = \langle \Delta_M (\phi_d(i)), \Delta_M (\phi_d(j)) \rangle_{\mathcal{H}} \), and \( E_{\phi_d}(i,j) = \langle \phi_d(i), \Delta_M (\phi_d(j)) \rangle_{\mathcal{H}}. \) Given the \( \mathbf{C}_d \)'s, the simple derivations show that the marginal product structure of the \( \zeta_m^{(i)} \)'s permits fast analytic computation of \( J_{\mathcal{M}}^* \) and \( R_{\mathcal{M}}^* \) based on the element-wise formulas

\[ J_{\mathcal{M}}^* (i,j) = \prod_{d=1}^D \bar{c}'_{d,i} J_{\phi_d} c_{d,j} \]

\[ R_{\mathcal{M}}^* (i,j) = \sum_{d=1}^D \left( \sum_{b \neq d} c'_{d,b} J_{\phi_d} c_{b,j} \right) \left( \bar{c}'_{d,i} E_{\phi_d} c_{d,j} \right) + \sum_{a,d \neq b,d \neq b} \left( \sum_{b \neq d} c'_{d,b} J_{\phi_d} c_{b,j} \right) \left( c'_{a,i} E_{\phi_d} c_{a,j} \right). \]

Notably, due to the marginal product structure of \( \zeta_m^{(i)} \), the \( D \)-dimensional integrals and partial derivatives required for the computation of \( J_{\mathcal{M}}^* \) and \( R_{\mathcal{M}}^* \) decompose into simple sums and products of integrals and partial derivatives over the marginal spaces. In contrast, computing such quantities for an arbitrary \( D \)-dimensional function is computationally prohibitive for moderately large \( D \). This highlights an important practical advantage of working with the marginal product structure: it facilitates efficient computation of \( D \)-dimensional integrals and partial derivatives which can serve as primitives for developing fast two-stage algorithms for more complex FDA procedures.

In practice, we form estimates \( \hat{C}_d \) using MARGARITA and then estimate the inner product matrices \( J_{\mathcal{M}}^* \) and \( R_{\mathcal{M}}^* \) by plugging \( \hat{C}_d \) into \( \text{(23)} \) and \( \text{(24)} \), respectively. Standard FDA techniques for rank and penalty parameter selection can be adopted to select \( K^\dagger \) and \( \lambda. \)

### 5. Simulation Study

#### 5.1. Representing Random Marginal Product Functions

In this section, we consider three methods for constructing the functional representation of a random sample generated from a marginal product functional model: (a) a TPB system estimated by the sandwich smoother (Xiao, Li, and Ruppert 2013), (b) the FCP-TPA algorithm (Allen 2013), and (c) the \( K \)-oMPB estimated using MARGARITA. The two competitors are widely used for multidimensional function representation, see Section S4 of the supplementary materials for more details.

The random function in our simulation is defined by the marginal product form: \( U(x) = \sum_{k=1}^N A_k^0 \prod_{d=1}^D \left( c'_{d,k} \phi_j(x_d) \right) \).

Here \( \phi_j \) is the period-1 Fourier basis, \( c'_{d,k} \) is the \( k \)th column vector of \( C_d^0 \), the fixed marginal factor matrix such that each element is an iid sample from \( N(0, 0.3^2) \); and \( A_k^0, \ldots, A_K^0 \sim N(0, \mathbf{C}_d^0). \)

The covariance matrix is constructed as \( \mathbf{C}_d^0 = \mathbf{ODO}' \), where \( O \) is a random \( K^\dagger \times K' \) orthogonal matrix and \( D \) is a diagonal matrix with \( D_{kk} = \exp(-0.7k^2) \) for \( k = 1, \ldots, K'. \) We took the function domain to be the unit cube, \( \mathcal{M} = [0, 1]^3 \). We fixed the true marginal basis dimensions to be \( m^d_{\text{true}} = 11 \) for all \( d \) and considered true ranks \( K_t = 10 \) and 20.

For both ranks, all combinations of the following sampling settings are considered. High versus low SNR; obtained by taking \( \sigma^2 \) to be 0.5 or 10, small versus large domain sample size; \( n_d = 30 \) or 50 for all \( d \), respectively, and small versus large subject sample size; where \( N \) is taken to be 5 or 50, respectively. For each of these settings, 100 replications are simulated according to Model (7). The performance of the fitting methods is assessed by computing the mean integrated
squared error (MISE) for each replication $r$: \[
\text{MISE}^{(r)} = \sum_{i=1}^{N^{(r)}} \left[ U^{(r)}_i(x) - \hat{U}^{(r)}_i(x) \right]^2 \text{dx}, \]
where $\hat{U}^{(r)}_i$ is an estimate of $U^{(r)}_i$ from the $r$th simulated dataset. Denote the Monte Carlo average of the MISE as \[
\text{moMISE} = 100^{-1} \sum_{r=1}^{100} \text{MISE}^{(r)}.
\]

For fitting, the second order derivative was used to define the marginal roughness penalties and a ridge penalty was used for regularization on the coefficients. Cubic b-splines were used as the marginal basis system.

We begin by investigating the performance as a function of rank for each combination of $m_d \in \{15, 25\}$ and $K_{\text{fit}} \in \{8, 15, 25\}$. A fair comparison between the TPB and MARGARITA should be based on enforcing (roughly) equivalently sized parameter spaces, that is, total number of degrees of freedom, so for TPB we take the smallest integer $m_d^{(\text{TPB})}$ such that $\prod_{d=1}^{D} m_d^{(\text{TPB})} \geq K_{\text{fit}} \sum_{d=1}^{D} m_d$ for comparison. To isolate the effects of the ranks, for each simulated dataset the models are estimated over a grid of smoothing parameters and the performance of the model with the lowest MISE is recorded. Section S5 in the supplementary materials presents a comprehensive comparison of the moMISE for each simulation setting and model parameterization. The results demonstrate that MARGARITA outperforms its competitors consistently, particularly in comparison to TPB fits that have similar degrees of freedom.

Figure 1 presents a comparison of FCP-TPA and MARGARITA for each combination of $m_d$ and $K_{\text{fit}}$, for several combinations of $N$ and $n_d$, with $K_f = 20$ and $\sigma^2 = 10$. While we see that in all cases MARGARITA results in fits with lower moMISE than FCP-TPA, we also observe that the ranks of the model have a significant impact on the performance. In practical settings, it is often necessary to automate the selection of these ranks as well as the smoothing parameters. Therefore, we compared the automated hyperparameter selection strategies for MARGARITA to the competitors automated smoothing approaches. Specifically, for the TPB method, we selected the smoothing parameters by minimizing the GCV criterion from Xiao, Li, and Ruppert (2013). For FCP-TPA, we implemented a $D$-dimensional extension of the nested cross-validation method from Huang, Shen, and Buja (2009), as suggested by the authors in Allen (2013). For our method, we selected penalty parameters using the cross-validation scheme outlined in Algorithm 2 of the supplementary text. To focus our analysis, we consider the large domain large sample case ($n_d = 50, N = 50$), with true rank $K_f = 20$ for both low and high SNRs ($\sigma^2 = 10, \sigma^2 = 0.5$) for 100 replications.

We use an elbow criteria to select the marginal ranks and set a threshold of $\text{PVG}(K) \geq 99.5\%$ for global rank selection. The Monte Carlo averages and standard error of these quantities are plotted for a range of $m$ and $K$ in the top right panels of Figures S1 and S2, respectively, found in the supplemental materials. We consistently identify a clear elbow at $\text{PVM}(m) = 15$ for both SNRs, which is in line with the results in Figure 1 showing a significant increase in performance for $m_d = 15$ compared to $m_d = 8$, while the performance boost from $m_d = 15$ to $m_d = 25$ is less pronounced. A $K_{\text{fit}} = 25$ is consistently selected across simulations. These ranks are fixed for subsequent comparison of the performance of smoothing parameter selec-

![Figure 1. MISE of the fits resulting from both FCP-TPA (gray) and MARGARITA (white). moMISE is denoted by a triangle. The Y-axis is plotted on log-scale for clarity.](image-url)
tion. Table 1 records the moMISE and accompanying standard errors for all methods, showing that MARGARITA’s automatic hyperparameter augmentation outperforms the competitors and is robust to noise. Furthermore, comparing these results to the corresponding results in the bottom right panel of Figure 1, we observe that our automated hyperparameter selection estimates models with similar performance to the ones obtained by selecting the oracle best fits over the hyperparameter grid.

Due to the super high-dimensional settings encountered, computational efficiency is as important a consideration as estimation performance in multidimensional FDA. Figure S4 in the supplemental text compares the computational time of FCP-TPA and MARGARITA for different simulation settings. We find that while both algorithms are comparable in computational speed for small \( N \) and small \( n_d \), MARGARITA outperforms FCP-TPA as \( N \) and \( n_d \) increase. This trend is expected, since increasing \( n_d \) does not increase the dimension of the optimization problem (10), while the factors estimated with FCP-TPA are of dimension \( n_d \), and thus the computational performance of the method can be expected to degrade as \( n_d \) increases and ultimately become infeasible in the fine grid limit.

### 5.2. Generalization Performance

In this section, we consider the generalization performance of MARGARITA, that is, how efficiently the \( K \)-OMPB estimated from a training sample of size \( N_{\text{train}} \) represents new realizations from the same distribution. The results of Section 2 indicate that we can expect near optimal generalization performance, with an inefficiency due to a “separability cost” that vanishes for increasing \( K \). We compare our a method to the marginal product FPCA procedure proposed in Chen, Delicado, and Müller (2017), referred to here as MARGFPCA, which provides a similar near optimality result. In brief, MARGFPCA constructs the marginal basis functions by applying FPCA to smoothed estimates of the marginal covariance functions.

The development of MARGFPCA focuses on the \( D = 2 \) case, so in this study we let the functional domain be \( \mathcal{M} = [0, 1]^2 \) and evaluate the generalization error of both MARGARITA and MARGFPCA as a function of \( N_{\text{train}} \) and \( K_{\text{fit}} \). We define random function \( U \) to be a nonstationary, nonseparable anisotropic Gaussian process which is observed over an equispaced 200 × 200 grid on \( \mathcal{M} \). For each combination of \( N_{\text{train}} \), and rank \( K_{\text{fit}} \), both MARGARITA and MARGFPCA are used to construct the representations for each of 50 realizations from an independent test set using least squares basis expansion. Each experimental set-up is repeated for 25 replications. Additional details on the definition of \( U \) and other simulation settings can be found in Supplemental Section S5.2.

Figure 2 (left) displays the average MISE on the test set, that is, the generalization error, as a function of \( K \) for both MARGARITA (red) and MARGFPCA (blue). The dotted and solid lines correspond to \( N_{\text{train}} = 10 \) and \( N_{\text{train}} = 100 \), respectively. For both training sample sizes, we observe that our method both uniformly outperforms MARGFPCA for all ranks considered and displays much faster convergence in \( K \). The green line shows the generalization performance of the eigenfunctions estimated using the two-stage FPCA procedure outlined in Section 4 for \( N_{\text{train}} = 100 \), with an initial MARGARITA of rank 60. The performance is nearly identical with that of the true eigenfunctions (black). Table S4 and Figure S5 in the supplemental material evaluate the two-stage estimates of first three eigenfunctions, showing accurate recovery as \( N \) increases. The right plot of Figure 2 gives the average MISE as a function of \( N_{\text{train}} \) for several \( K \). Recalling that MARGARITA is only guaranteed to converge to a local solution, these result indicate that, at least in some cases, the local (computable) solution still exhibits good convergence.

### Table 1. Monte Carlo average MISE for the \( n_d = 50, N = 50, K_t = 20 \) regime for both high and low SNRs.

| Method  | FCP-TPA | MARGARITA | TPB |
|---------|---------|-----------|-----|
| High SNR | 0.0729 ± 0.0009 | 0.0418 ± 0.0004 | 0.5927 ± 0.0060 |
| Low SNR  | 0.0886 ± 0.0009 | 0.0458 ± 0.0004 | 0.6681 ± 0.0061 |

Note: Each method’s proposed automatic penalty parameter selection method was used for estimation.

![Figure 2](image-url)
properties. Results for more ranks and training sample sizes are recorded in Section S5 of the supplemental materials and yield similar conclusions.

6. Real Data Analysis

The white matter (WM) of the human brain consists of large collections of myelinated neural fibers that permit fast communication between different regions of the brain. Diffusion magnetic resonance imaging (dMRI) is a noninvasive imaging technique which uses spatially localized measurements of the diffusion of water molecules to probe the WM microstructure. At each three-dimensional voxel in the brain, the diffusion image can be used to compute scalar summaries of local diffusion, for example, fractional anisotropy (FA) or mean diffusivity. The resulting data can be organized as a mode-3 tensor. For this application, we consider a dataset consisting of the brain images of 50 subjects in an age matched balanced case-control traumatic brain injury (TBI) study. Previous studies have shown the potential for using FA to identify white matter abnormalities associated with TBI and post concussive syndrome (Kraus et al. 2007). Typically, voxel-based analysis are performed for group-wise analysis of FA using Tract-Based Spatial Statistics (TBSS) (Smith et al. 2006), though such analysis are often not able to establish significant group differences (Khong et al. 2016), partially due to low power resulting from the large voxel-based multiple testing problem. Due to the continuity of the diffusion process, the FA tensor can be considered as discrete noisy observations of an underlying multidimensional random field, hence, we may adopt the statistical model in (7). In this analysis, we focus on a functional approach to predict disease status and identify regions in the WM which differ significantly between TBI and control. For details on the study design, MRI scanning protocol, and dMRI preprocessing, please visit Section S6 in the supplementary material.

The voxel grid is of size 115 × 140 × 120. Point-wise estimates of the mean function at each voxel are obtained using the sample mean tensor, which is then used to center the data. Equispaced cubic b-splines of ranks \( m_1 = 57, m_2 = 70, m_3 = 60 \), selected using a 90% threshold on the quantity described in Section 3.5, are used as marginal basis systems. Marginal roughness is penalized by the second order derivative and coefficients were regularized with a ridge penalty, with penalty parameters \( \lambda_d = 10^{-10} \) for \( d = 1, 2, 3 \) and \( \lambda_4 = 10^{-8} \). A rank \( K = 500 \) model is estimated from the mean centered data tensor using MARGARITA. FPCA is then performed on the represented data using the fast two-stage approach outlined in Section 4. The first 45 eigenfunctions, denoted collectively as \( \psi \), explain \( \approx 99\% \) of the represented variance and are used in constructing the final continuous representations of data. A lasso penalized logistic regression classifier is trained to predict disease status using the subject coefficient vectors obtained by their representation over \( \psi \). The resulting classification performance is evaluated using leave-one-out cross validation (LOOCV). To localize group differences to particular eigenfunctions, univariate permutation test are performed on the coefficients and the resulting \( p \)-values are corrected to maintain a false discovery rate (FDR) \( \leq 5\% \) using Benjamini and Hochberg (1995). Finally, data-driven regions of interest (ROIs) are defined as spatial volumes where the values of the significant eigenfunctions are “extreme”, that is, outside the 0.5% and 99.5% quantiles.

The LOOCV accuracy, precision and recall are 0.96, 1.0, and 0.92, respectively, indicating substantial discriminatory power of the learned basis functions. Additionally, the testing procedure identified significant group differences in the coefficients of three eigenfunctions. For comparison, we applied TBSS to this data and no significant group differences were identified.

Figure 3 shows two cross sections of the brain, with the data-driven ROIs corresponding to the identified eigenfunctions displayed in blue, red, and green. The ROIs in Figure 3(a) are found within areas of the middle cerebellar peduncle (MCP) and, in Figure 3(b), in areas along the superior longitudinal fasciculus (SLF) fiber bundle. Wang et al. (2016) found increased FA in the MCP is associated with increased cognitive impairment. Xiong

Figure 3. Data-driven ROIs created from thresholding the 0.5% and 99.5% quantiles of the three identified eigenfunctions in blue, red, and green.
et al. (2014) found decreased FA in the SLF in patients with TBI. We note that both of these studies were completed in acute cases of TBI, whereas our data represents a more chronic state of TBI, often called post-concussive syndrome. That being said, these tracts are thought to be altered because of the nature of biophysical forces suffered in TBI. In all TBI, there is rotation of the head around the neck, which causes shearing and stretching of the brain stem tracts. In addition, the longer tracts in the brain, including the SLF, are subject to shearing forces on left to right rotation of the head around the neck. In fact, Post et al. (2013) found that mechanical strain in the brain stem and cerebellum are significantly correlated with angular acceleration of the brain, suggesting fibers in this area are susceptible to changes related to TBI. Therefore, our findings of changes in the MCP and SLF are consistent with the hypothesized mechanism and previous findings in TBI.

7. Discussion and Future Work

Our work introduces a methodological framework and accompanying estimation algorithm for constructing a flexible and efficient continuous representation of multidimensional functional data. We consider basis functions that exhibit a marginal product structure and prove that an optimal set of such functions can be defined by the penalized tensor decomposition of an appropriate transformation of the raw data tensor. A variety of separable roughness penalties can be used to promote smoothness. Regularized parameter estimation is performed using a block coordinate descent scheme and we describe globally convergent numerical algorithms for solving the subproblems. Using extensive simulation studies, we illustrate the superiority of our proposed method compared to competing alternatives. In a real data application of the group-wise analysis of diffusion MRI, we show that our method can facilitate the prediction of disease status and identify biologically meaningful ROIs.

This work can be extended in several interesting directions. A principled and computationally efficient approach to both a generalization of cross-validation criteria and model-based selection of the global rank, respectively, are of interest. Additionally, many modern functional datasets are observed irregularly over the domain, rather than the common grid we consider here. To use our method on dense irregular data, we can bin the data using a common grid and define the observed data tensor to be the bin-specific sample means for each subject. However, this approach is problematic for sparsely sampled irregular data, requiring further extension of the method.

Supplementary Materials

eMFDA_supp.pdf Supporting technical details for the method and pseudocode for algorithms, proofs for all theorems, additional simulation studies and results and real data description.

Acknowledgments

We extend our gratitude to the editor, associate editor, and two reviewers for their valuable feedback, which played a significant role in improving the article.

Disclosure Statement

The authors report no competing interest.

Funding

This work was partially supported by the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health.

ORCID

Xing Qiu https://orcid.org/0000-0002-2330-3544

References

Allen, G. I. (2013), "Multi-Way Functional Principal Components Analysis," in 2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 220–223. [568,572,573]
Allen, G. I., and Weylandt, M. (2019), "Sparse and Functional Principal Components Analysis," 2019 IEEE Data Science Workshop (DSW). [568]
Bartels, R. H., and Stewart, G. W. (1972), "Solution of the Matrix Equation AX + XB = C," Communications of the ACM, 15, 820–826. [571]
Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society, Series B, 57, 289–300. [575]
Bertsekas, D. P. (1997), "Nonlinear Programming," Journal of the Operational Research Society, 48, 334–334. [571]
Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," Foundations and Trends in Machine Learning, 3, 1–122. [571]
Chen, K., Delicado, P., and Müller, H.-G. (2017), "Modelling Function-Valued Stochastic Processes, with Applications to Fertility Dynamics," Journal of the Royal Statistical Society, Series B, 79, 177–196. [568,574]
Chen, L.-H., and Jiang, C.-R. (2017), "Multi-Dimensional Functional Principal Component Analysis," Statistics and Computing, 27, 1181–1192. [567]
Happ, C., and Greven, S. (2018), "Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains," Journal of the American Statistical Association, 113, 649–659. [568]
Hilliar, C. J., and Lim, L.-H. (2013), "Most Tensor Problems Are NP-Hard," Journal of the ACM, 60, 1–39. [570]
Hsing, T., and Eubank, R. (2015), Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, Chichester: Wiley. [567,568]
Huang, J. Z., Shen, H., and Buja, A. (2009), "The Analysis of Two-Way Functional Data using Two-Way Regularized Singular Value Decompositions," Journal of the American Statistical Association, 104, 1609–1620. [568,573]
Huang, K., Sidiropoulos, N. D., and Liavas, A. P. (2016), "A Flexible and Efficient Algorithmic Framework for Constrained Matrix and Tensor Factorization," IEEE Transactions on Signal Processing, 64, 5052–5065. [571]
Khong, E., Odenwald, N., Hashim, E., and Cusimano, M. D. (2016), "Diffusion Tensor Imaging Findings in Post-Concussion Syndrome Patients After Mild Traumatic Brain Injury: A Systematic Review," Frontiers in Neurology, 7, 156. [575]
Kraus, M. E., Susmaraas, T., Caughlin, B. P., Walker, C. J., Sweeney, J. A., and Little, D. M. (2007), "White Matter Integirity and Cognition in Chronic Traumatic Brain Injury: A Diffusion Tensor Imaging Study," Brain, 130, 2508–2519. [575]
Li, C., Xiao, L., and Luo, S. (2020), "Fast Covariance Estimation for Multivariate Sparse Functional Data," Stat, 9, e245. [568]
Li, Y., Huang, C., and Härdele, W. K. (2019), "Spatial Functional Principal Component Analysis with Applications to Brain Image Data," Journal of Multivariate Analysis, 170, 263–274. Special Issue on Functional Data Analysis and Related Topics. [567]
Lynch, B., and Chen, K. (2018), “A Test of Weak Separability for Multi-Way Functional Data, with Application to Brain Connectivity Studies,” *Biometrika*, 105, 815–831. [568]

Phan, A., Tichavský, P., and Cichocki, A. (2013), “Fast Alternating Is Algorithms for High Order candeomp/parafac Tensor Factorizations,” *IEEE Transactions on Signal Processing*, 61, 4834–4846. [571]

Post, A., Oeur, A., Hoshizaki, B., and Gilchrist, M. D. (2013), “Examination of the Relationship between Peak Linear and Angular Accelerations to Brain Deformation Metrics in Hockey Helmet Impacts,” *Computer Methods in Biomechanics and Biomedical Engineering*, 16, 511–519. [576]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, New York: Springer-Verlag. [567,572]

Silverman, B. W. (1996), “Smoothed Functional Principal Components Analysis by Choice of Norm,” *The Annals of Statistics*, 24, 1–24. [567,571]

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M., and Behrens, T. E. (2006), “Tract-based Spatial Statistics: Voxelwise Analysis of Multi-Subject Diffusion Data,” *NeuroImage*, 31, 1487–1505. [575]

Stone, C. J. (1980), “Optimal Rates of Convergence for Nonparametric Estimators,” *The Annals of Statistics*, 8, 1348–1360. [567]

Wang, J., Wong, R. K. W., and Zhang, X. (2020), “Low-Rank Covariance Function Estimation for Multidimensional Functional Data,” *Journal of the American Statistical Association*, 117, 809–822. [567]

Wang, Z., Wu, W., Liu, Y., Wang, T., Chen, X., Zhang, J., Zhou, G., and Chen, R. (2016), “Altered Cerebellar White Matter Integrity in Patients with Mild Traumatic Brain Injury in the Acute Stage,” *PLoS One*, 11, e0151489. [575]

Wasserman, L. (2010), *All of Nonparametric Statistics*, New York: Springer. [567]

Xiao, L., Li, Y., and Ruppert, D. (2013), “Fast Bivariate p-splines: The Sandwich Smoother,” *Journal of the Royal Statistical Society*, Series B, 75, 577–599. [572,573]

Xiong, K., Zhu, Y., Zhang, Y., Yin, Z., Zhang, J., Qiu, M., and Zhang W. (2014), “White Matter Integrity and Cognition in Mild Traumatic Brain Injury Following Motor Vehicle Accident,” *Brain Research*, 1591, 86–92. [576]

Zhang, J.-T., and Chen, J. (2007), “Statistical Inferences for Functional Data,” *The Annals of Statistics*, 35, 1052–1079. [567]