Efficient Multidimensional Functional Data Analysis Using Marginal Product Basis Systems

Journal:	*Journal of the Royal Statistical Society: Series B*
Manuscript ID	JRSSB-Dec-2021-0641
Manuscript Type:	Original Article
Date Submitted by the Author:	13-Dec-2021
Complete List of Authors:	Consagra, William; University of Rochester, Department of Biostatistics and Computational Biology Venkataraman, Arun; University of Rochester, Department of Physics and Astronomy Qiu, Xing; University of Rochester Medical Center, Biostatistics and Computational Biology

ORIGINAL ARTICLE

Efficient Multidimensional Functional Data Analysis Using Marginal Product Basis Systems

William Consagra[1] | **Arun Venkataraman**[2] | **Xing Qiu**[1]

[1]Department of Biostatistics and Computational Biology , University of Rochester, U.S.A.

[2]Department of Physics and Astronomy, University of Rochester, U.S.A.

Correspondence
Xing Qiu, Department of Biostatistics and Computational Biology, University of Rochester, U.S.A. 14642
Email: xing_qiu@urmc.rochester.edu

Funding information
This work is supported in part by the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health.

Many modern datasets, from areas such as neuroimaging and geostatistics, come in the form of a random sample of tensor-valued data which can be understood as noisy observations of a smooth multidimensional random function. Most of the traditional techniques from functional data analysis are plagued by the curse of dimensionality and quickly become intractable as the dimension of the domain increases. In this paper, we propose a framework for learning continuous representations from a sample of multidimensional functional data that is immune to several manifestations of the curse. These 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 established. The advantages of our method over competing methods are demonstrated in a simulation study. We conclude with a real data application in neuroimaging.

Keywords: Basis representation; tensor decomposition; universal approximation; functional principal component analysis

1 | INTRODUCTION

Functional data analysis (FDA) is a subfield of statistics concerned with the analysis of collections of smooth functions. Most of the foundational work in FDA considers the 1-dimensional case, i.e. when the function domain of definition is an interval on $\mathbb{R}^1$, usually understood to represent time (Ramsay and Silverman, 2005; Hsing and Eubank, 2015). Although many modern datasets from fields such as neuroimaging, chemometrics, climate science, and astronomy can be modeled as random functions/fields defined on multidimensional and/or non-Euclidean domains, there has been a relative lack of attention to extending FDA techniques to these cases, mainly due to the so-called *curse of dimensionality* that makes the traditional FDA approaches computationally intractable.

In most FDA applications, the analyst needs to perform the initial step of estimating a smooth function from each subject's discretely observed, noisy data (Zhang and Chen, 2007), which we refer to here as *functional representation*. A host of downstream analyses, such as functional principal components analysis (FPCA), regression, and differential equation modeling, are then performed on these reconstructed smooth functions. In 1-dimension, functional representation is typically accomplished by nonparametric estimation through either local polynomial/kernel regression or by expansion over some appropriately defined basis system, e.g. various splines. Unfortunately, these approaches suffer from the curse of dimensionality when extended to multidimensional cases. 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, the dimension of the domain (Wasserman, 2010). To avoid these issues, a common tactic is to impose some structural assumptions on the nature of the underlying functions, leading to the development of the so-called semiparametric regression models such as additive and single index models (Ruppert et al., 2006). Although the semi-structured nature permits efficient estimation, these frameworks are often overly restrictive for real word data exhibiting complex dependence patterns.

In what follows, we propose a framework for multidimensional FDA based on learning the optimal *marginal product basis* (MPB) for representing realizations of a D-dimensional random field. Heuristically, a function is referred to as a marginal product function (MPF) if it is multiplicatively separable over some product domain, see Definition 2.1; and an MPB is simply a collection of linearly independent MPFs. Crucially, the number of parameters needed to estimate the MPB is not exponential in D, yet it still retains much of the flexibility of the nonparametric models. This structure has been used to facilitate efficient procedures for a variety of related tasks in multidimensional function approximation from scattered data (Beylkin et al., 2009; Chevreuil et al., 2015; Suzuki et al., 2016; Kargas and Sidiropoulos, 2021), or data observed on a grid (Grasedyck et al., 2013; Gorodetsky et al., 2019), for estimating the fixed effects in functional ANOVA (Huang et al., 2009) and in the context of reduced basis methods and dictionary learning (Nouy, 2017).

Owing to well known optimality properties, much of the attention toward designing optimal basis systems for representing functional data has focused on estimating the eigenfunctions of the covariance operator (Silverman, 1996; Yao et al., 2005). Estimating the eigenfunctions, i.e. performing FPCA, for a general D-dimensional random function requires the estimation of the $2D$-dimensional covariance function, denoted $C(\boldsymbol{x}, \boldsymbol{y})$. Some techniques have recently been proposed (Chen and Jiang, 2017; Li et al., 2019; Wang et al., 2020), although this approach can become untenable as D becomes even moderately large. To alleviate the computational difficulties associated with estimating a generic covariance function, a common tactic in the literature has been to assume some notion of separability for C, hence reducing the computations to the marginal covariance kernels. In Chen et al. (2017), the authors propose a marginal product FPCA for $D = 2$ and Lynch and Chen (2018) show that this marginal product FPCA is optimal under weak separability on C. Testing procedures for assessing separability in functional data have also been proposed (Aston et al., 2017; Liang et al., 2021). Masak and Panaretos (2019) consider estimation of C under additive perturbation of separability and then, in a follow-up work for the special case of $D = 2$, under a general non-

parameteric structure of C using a separable components decomposition (Masak et al., 2020). Notions of separable covariance structures have also been used in related work for estimation of functional graphical models (Zapata et al., 2020). The aforementioned works are mostly developed for the $D = 2$ case, though for many, extending the theory to $D > 2$ is relatively straightforward. Computationally speaking, it is a different story. For instance, when $D > 2$, the marginal product FPCA in Chen et al. (2017) requires multidimensional numerical integration or non-parametric smoothing in order to estimate the marginal covariance functions, re-introducing a manifestation of the curse of dimensionality. In this work, we show that by using the MPB structure, we can estimate a data-adaptive function space for efficient representation of functional data that is immune to such manifestations of the curse. At the heart of our method is the identification of an isometric embedding which allows us to reparameterize our problem into to a lower dimensional space whose dimension is user controlled. This permits the derivation of fast approximation algorithms which scale favorably with huge datasets.

The rest of the paper is organized as follows. In Section 2, we formulate the optimal MPB system and then derive an efficient estimation procedure based on the canonical polyadic decomposition (CPD) of $\widehat{\mathcal{G}}$, the reduced data tensor. Incorporation of roughness-based regularization using differential operators is also discussed in this section. Relevant theoretical properties are established in Section 3. Section 4 compares the proposed method with competing methods in simulation studies. In Section 5, we analyze a set of brain imaging data collected from subjects who suffered from traumatic brain injury. Section 6 offers concluding remarks and potential future directions.

2 | MODEL AND METHODS

2.1 | Background and Model Description

In this study, we are interested in modeling multidimensional random functions U with real-valued square integrable realizations u, i.e. $u(\boldsymbol{x}) \sim U \in \mathcal{H} := \mathbb{L}^2(\mathcal{M})$. Here $\boldsymbol{x} = (x_1, \ldots, x_D)' \in \mathcal{M}$; each x_d, for $d = 1, \ldots, D$, is a member in the marginal domain $\mathcal{M}_d$, which is assumed to be a compact subset of Euclidean space $\mathbb{R}^{p_d}$; and the joint domain of $u(\boldsymbol{x})$ can be decomposed as $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_D$. We make the following regularity assumptions on U:

Assumption 1. Assume that: (a) $\mathbb{E}[U] = 0$, (b) $\mathbb{E}[\int_{\mathcal{M}} U^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}] < \infty$, and (c) U is *mean-square continuous*, that is, for any $\boldsymbol{x} \in \mathcal{M}$ and any sequence $\{\boldsymbol{x}_n\}$ in $\mathcal{M}$ converging to $\boldsymbol{x}$, then

$$\lim_{n\to\infty} \mathbb{E}\left[(U(\boldsymbol{x}_n) - U(\boldsymbol{x}))^2\right] = 0.$$

The mean zero assumption on U is made for convenience of presentation and would not fundamentally change any of our results. The mean square integrability and mean-square continuity assumptions are a standard requirement (Hsing and Eubank, 2015).

Let $H_d := \mathbb{L}^2(\mathcal{M}_d)$, so that $\mathcal{H} := \mathbb{L}^2(\mathcal{M}) = \bigotimes_{d=1}^{D} H_d$, the tensor product of D member spaces. We assume that there is a pre-defined complete basis system, $\boldsymbol{\phi}_d := \{\boldsymbol{\phi}_{d,j}\}_{j=1}^{\infty}$, for each marginal functional space H_d. Denote their rank-m_d truncations as $\boldsymbol{\phi}_{m_d,d} = (\boldsymbol{\phi}_{d,1}, ..., \boldsymbol{\phi}_{d,m_d})'$; $H_{m_d,d} := \mathrm{span}(\boldsymbol{\phi}_{m_d,d})$; and $\mathcal{H}_{\boldsymbol{m}} := \bigotimes_{d=1}^{D} H_{m_d,d}$. Here $\boldsymbol{m} = (m_1, \ldots, m_D)'$ determines the expressiveness of the truncated basis systems of the member spaces.

By construction,

$$\boldsymbol{\tau} := \bigotimes_{d=1}^{D} \boldsymbol{\phi}_d = \left\{\tau_{j_1,\ldots,j_D}(\boldsymbol{x}) = \prod_{d=1}^{D} \phi_{d,j_d}(x_d), j_d = 1,\ldots,\infty\right\},$$
$$\boldsymbol{\tau}_{\boldsymbol{m}} := \bigotimes_{d=1}^{D} \boldsymbol{\phi}_{m_d,d} = \left\{\tau_{j_1,\ldots,j_D}(\boldsymbol{x}) = \prod_{d=1}^{D} \phi_{d,j_d}(x_d), j_d = 1,\ldots,m_d\right\} \tag{1}$$

are the complete tensor product bases (TPB) for $\mathcal{H}$ and $\mathcal{H}_{\boldsymbol{m}}$, respectively. Let $u = \sum_{j_1,\ldots,j_D} a_{j_1,\ldots,j_D}\phi_{j_1,\ldots,j_D}$ and $v = \sum_{j'_1,\ldots,j'_D} b_{j'_1,\ldots,j'_D}\phi_{j'_1,\ldots,j'_D}$ be two elements in $\mathcal{H}$. Their inner product can be represented as follows

$$\langle u,\ w\rangle_{\mathcal{H}} := \sum_{j_1,\ldots,j_D}\sum_{j'_1,\ldots,j'_D}\left[a_{j_1,\ldots,j_D} b_{j'_1,\ldots,j'_D}\prod_{d=1}^{D}\langle\phi_{d,j_d},\ \phi_{d,j'_d}\rangle_{H_d}\right]. \tag{2}$$

Definition 2.1 (Marginal product structure). $\zeta \in \mathcal{H}$ is called a rank-1 marginal product function (MPF), or simply an MPF if there is no ambiguity, if it is *multiplicatively separable*:

$$\zeta(\boldsymbol{x}) = \prod_{d=1}^{D}\xi_d(x_d), \qquad \xi_d \in H_d. \tag{3}$$

As an extension, $u(\boldsymbol{x}) \in \mathcal{H}$ is called a rank-K marginal product function (K-MPF), or it has a rank-K *marginal product structure*, if it is a linear combination of K linearly independent rank-1 MPFs

$$u(\boldsymbol{x}) = \sum_{k=1}^{K} b_k\zeta_k(\mathbf{x}) = \sum_{k=1}^{K} b_k\prod_{d=1}^{D}\xi_{k,d}(x_d), \qquad \xi_{k,d} \in H_d, \quad b_k \in \mathbb{R}. \tag{4}$$

We denote the collections of rank-1 MPFs for both the infinite and finite dimensional spaces

$$\mathcal{L} := \left\{\zeta(\boldsymbol{x}) : \zeta(\boldsymbol{x}) = \prod_{d=1}^{D}\xi_d(x_d),\ \xi_d \in H_d,\ \|\xi_d\|_{H_d} = 1\right\}$$
$$\mathcal{L}_{\boldsymbol{m}} := \left\{\zeta(\boldsymbol{x}) : \zeta(\boldsymbol{x}) = \prod_{d=1}^{D}\xi_d(x_d),\ \xi_d \in H_{m_d,d},\ \|\xi_d\|_{H_d} = 1\right\} \tag{5}$$

Note that the unit norm condition $\|\xi_d\|_{H_d} = 1$ is needed to resolve scale identifiability issues. In this work, we propose to estimate the optimal basis of K-MPFs for representing realizations of U, a notion formalized as follows:

Definition 2.2 (Optimal Rank-K MPB). Denote the set of linearly independent K-MPFs

$$\mathcal{V}_K := \left\{\boldsymbol{\zeta} = (\zeta_1,\ldots,\zeta_K)' : \zeta_k \in \mathcal{L}, \zeta_1,\ldots,\zeta_K \text{ linearly independent}\right\}. \tag{6}$$

The optimal rank-K MPB, denoted K-oMPB, is defined as the solution to

$$\boldsymbol{\zeta}^* = \underset{\boldsymbol{\zeta}\in\mathcal{V}_K}{\arg\inf} \quad \mathbb{E}\|U - P_{\boldsymbol{\zeta}}(U)\|^2_{\mathcal{H}} \tag{7}$$

where $P_{\boldsymbol{\zeta}}$ is the projection operator onto span($\boldsymbol{\zeta}$).

Equation (7) is an optimization problem over the infinite dimensional space $\mathcal{V}_K$ and therefore additional structure is necessary. Define the space of functions

$$\mathcal{V}_{K,\boldsymbol{m}} := \left\{\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_K)' : \zeta_k \in \mathcal{L}_{\boldsymbol{m}}, \zeta_1, ..., \zeta_K \text{ linearly independent}\right\}, \tag{8}$$

and define the associated best rank K-oMPB

$$\boldsymbol{\zeta}^*_{\boldsymbol{m}} = \underset{\boldsymbol{\zeta}\in\mathcal{V}_{K,\boldsymbol{m}}}{\arg\inf} \quad \mathbb{E}\|U - P_{\boldsymbol{\zeta}}(U)\|^2_{\mathcal{H}}. \tag{9}$$

Given a random sample of N realizations $U_i \sim U$, define the corresponding empirical estimate of (7) as

$$\check{\boldsymbol{\zeta}}^*_{\boldsymbol{m},N} = \underset{\boldsymbol{\zeta}\in\mathcal{V}_{K,\boldsymbol{m}}}{\arg\inf} \quad \frac{1}{N}\sum_{i=1}^{N}\|U_i - P_{\boldsymbol{\zeta}}(U_i)\|^2_{\mathcal{H}} \tag{10}$$

In practical applications, the U_i are observed with noise at each discrete location in $\mathcal{X} \subset \mathcal{M}$, where

$$\mathcal{X} = (x_{11}, x_{12}, ..., x_{1n_1})' \times (x_{21}, x_{22}, ..., x_{2n_2})' \times \cdots \times (x_{D1}, x_{D2}, ..., x_{Dn_D})'$$

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

$$\begin{aligned}\mathcal{Y}(i_1, i_2, ..., i_D, i) = U_i(x_{1,i_1}, x_{2,i_2}, ..., x_{D,i_D}) + \mathcal{E}(i_1, i_2, ..., i_D, i)\\ \text{for } i_d = 1, 2, ..., n_d,\ d = 1, 2, ..., D,\ i = 1, 2, ..., N\end{aligned} \tag{11}$$

where $\mathcal{Y}$ is a $D+1$-mode tensor with dimensions $(n_1, n_2, ..., n_D, N)$, $\mathbb{E}[\text{vec}(\mathcal{E})] = \boldsymbol{0}$ and $\text{Var}[\text{vec}(\mathcal{E})] = \sigma^2\boldsymbol{I}$. The discretized counterpart to (10) is given by

$$\widehat{\boldsymbol{\zeta}}^*_{N,\boldsymbol{m}} := \underset{\boldsymbol{\zeta}\in\mathcal{V}_{K,\boldsymbol{m}}}{\arg\inf} \min_{\boldsymbol{B}\in\mathbb{R}^{N\times K}} \quad \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{Y}_i - \sum_{k=1}^{K}\boldsymbol{B}_{i,k}\bigotimes_{d=1}^{D}\boldsymbol{\xi}_{d,k}\right\|^2_F. \tag{12}$$

where $\mathcal{Y}_i \in \mathbb{R}^{n_1\times\cdots\times n_D}$ is observed data tensor for the ith realization, $\boldsymbol{\xi}_{d,k} \in \mathbb{R}^{n_d}$ is the evaluation of $\xi_{k,d}$ on $\boldsymbol{x}_d$ and $\boldsymbol{B}$ is the matrix of coefficients for the ζ_k's for each of the N samples.

2.2 | Estimation

One approach to solving (12) is to estimate the $\boldsymbol{\xi}_{d,k}$ using the (regularized) tensor decomposition of $\mathcal{Y}$, e.g. by the FCP-TPA method developed in Allen (2013), and then perform a basis expansion over $\phi_{m_d,d}$. We refer to this as the decompose then represent approach. Unfortunately, this approach scales poorly for large tensors, as the number of unknown parameters increases unboundedly as $\mathcal{X}$ becomes arbitrarily dense. In this section, we propose an alternative approach which is able to both estimate the basis functions directly and control the dimension of the optimization problem.

2.2.1 | A Convenient Reparameterization

First we note that for $\zeta \in \mathcal{V}_{K,\boldsymbol{m}}$, the marginal basis function can be written as $\xi_{k,d}(x_d) = \sum_{j=1}^{m_d} c_{d,k,j}\phi_{d,j}(x_d)$. Consequently $\widehat{\zeta}^*_{N,\boldsymbol{m}}$ is equivalently defined by the solutions to the following optimization problem

$$(\widehat{\boldsymbol{C}}_1, ..., \widehat{\boldsymbol{C}}_D) := \underset{(\boldsymbol{C}_1,...,\boldsymbol{C}_D)\in\mathcal{C}_{K,\boldsymbol{m}}}{\arg\inf} \min_{\boldsymbol{B}\in\mathbb{R}^{N\times K}} \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{y}_i - \sum_{k=1}^{K}\boldsymbol{B}_{i,k}\bigotimes_{d=1}^{D}\boldsymbol{\Phi}_d\boldsymbol{c}_{d,k}\right\|_F^2 \tag{13}$$

where $\boldsymbol{\Phi}_d \in \mathbb{R}^{n_d\times m_d}$ is the evaluation of ϕ_d on the marginal grid $\boldsymbol{x}_d$, i.e. $\boldsymbol{\Phi}_{d,i_d j_d} := \phi_{d,j_d}(x_{d,i_d})$, $\boldsymbol{c}_{d,k} \in \mathbb{R}^{m_d}$ is the coefficient of $\xi_{k,d}$, and $\boldsymbol{C}_d$ is the matrix whose columns are the $\boldsymbol{c}_{d,k}$. Let $\boldsymbol{J}_{\phi_d}$ be the matrix of pairwise $\mathcal{H}_{m,d}$ inner products of $\phi_{m,d}$. The reparameterization of $\mathcal{V}_{K,\boldsymbol{m}}$ is defined as

$$\mathcal{C}_{K,\boldsymbol{m}} = \{(\boldsymbol{C}_1, ..., \boldsymbol{C}_D) : \boldsymbol{c}'_{d,k}\boldsymbol{J}_{\phi_d}\boldsymbol{c}_{d,k} = 1 \text{ for } d = 1, ..., D; k = 1, ..., K\}$$

Denote the SVD of the basis evaluation matrices $\boldsymbol{\Phi}_d = \boldsymbol{U}_d\boldsymbol{D}_d\boldsymbol{V}'_d$. In general, we have $n_d > m_d$, so $\boldsymbol{U}_d \in \mathbb{R}^{n_d\times m_d}$ is a semi-orthogonal matrix; $\boldsymbol{D}_d \in \mathbb{R}^{m_d\times m_d}$ is an invertible diagonal matrix; and $\boldsymbol{V}_d \in \mathbb{R}^{m_d\times m_d}$ is an orthogonal matrix. For any $\zeta \in \mathcal{V}_{K,\boldsymbol{m}}$, the evaluation of the MPB functions $\boldsymbol{\xi}_d = (\xi_{d1}, ..., \xi_{dK})'$ on $\boldsymbol{x}_d$, is represented as

$$\Xi_d = \boldsymbol{\Phi}_d\boldsymbol{C}_d = \boldsymbol{U}_d\boldsymbol{D}_d\boldsymbol{V}'_d\boldsymbol{C}_d = \boldsymbol{U}_d\tilde{\boldsymbol{C}}_d, \qquad \tilde{\boldsymbol{C}}_d := \boldsymbol{D}_d\boldsymbol{V}'_d\boldsymbol{C}_d.$$

The following theorem proves the equivalence between the solution of Equation (13) and the rank-K CPD of an appropriately defined tensor.

Theorem 2.1 (Functional Tensor Decomposition Theorem). *Define $\widehat{\mathcal{G}} := \mathcal{Y} \times_1 \boldsymbol{U}'_1 \times_2 \boldsymbol{U}'_2 \cdots \times_D \boldsymbol{U}'_D$, which is a $(D+1)$-mode tensor with dimensions $(m_1, m_2, ..., m_D, N)$ and denote its rank-K decomposition by $\widehat{\mathcal{G}}_K(\check{\boldsymbol{B}}, \check{\boldsymbol{C}}) = \sum_{k=1}^{K}\left[\bigotimes_{d=1}^{D}\check{\boldsymbol{c}}_{d,k}\right]\otimes\check{\boldsymbol{b}}_k$, with factor matrices $\check{\boldsymbol{B}} \in \mathbb{R}^{N\times K}$ and $\check{\boldsymbol{C}} = [\check{\boldsymbol{C}}_1, \ldots, \check{\boldsymbol{C}}_D]$, $\check{\boldsymbol{C}}_d \in \mathbb{R}^{m_d\times K}$. $\check{\boldsymbol{c}}_{d,k}$ and $\check{\boldsymbol{b}}_k$ are the kth column of $\check{\boldsymbol{C}}_d$ and $\check{\boldsymbol{B}}$, respectively. The optimization problem* (13) *has the following solutions: $\widehat{\boldsymbol{B}} = \check{\boldsymbol{B}}$ and $\widehat{\boldsymbol{C}}_d = \boldsymbol{V}_d\boldsymbol{D}_d^{-1}\check{\boldsymbol{C}}_d$, for $d = 1, \ldots, D$.*

Proof. Refer to Section S1 of the Supplemental Materials. □

As a remark, "$\times_d$" is called d-mode multiplication in tensor algebra. See e.g. Kolda and Bader (2009) for more details about this operation. Theorem 2.1 shows that estimating the K-oMPB is equivalent to the CPD of the $\widehat{\mathcal{G}}$ tensor. As the dimensionality of $\widehat{\mathcal{G}}$ is controlled by the rank of $\phi_{m_d,d}$, as opposed to the number of marginal grid points n_d, this defines a practical reduction transformation which permits user control of the dimensionality of the optimization problem.

2.2.2 | Regularization

In order to ameliorate the influence of noise and promote smoothness in the estimated K-oMPB basis, it is desirable to incorporate a non-negative penalty functional to the objective function in Equation (13). We consider penalties of the following form:

Definition 2.3 (Separable Roughness Penalty Functional). Denote the Sobolev space over the dth marginal domain $\mathbb{W}^{\alpha_d,2}(\mathcal{M}_d)$ for some order α_d. Let $u_d(x_d) \in \mathbb{W}^{\alpha_d,2}(\mathcal{M}_d)$, and $u(\mathbf{x}) = u_1(x_1)\cdots u_D(x_D) \in \mathbb{W}^{\alpha_1,2}(\mathcal{M}_1)\times\cdots\times\mathbb{W}^{\alpha_D,2}(\mathcal{M}_D)$.

We assume that the penalty functional Pen : $\mathbb{W}^{\alpha_1,2}(\mathcal{M}_1) \times \cdots \times \mathbb{W}^{\alpha_D,2}(\mathcal{M}_D) \to [0, \infty)$ can be represented as

$$\text{Pen}(u(\mathbf{x})) = \int_{\mathcal{M}} \sum_{d=1}^{D} \lambda_d L_d^2(u_d)$$

for some $\lambda_d > 0$, where $L_d : \mathbb{W}^{\alpha_d,2}(\mathcal{M}_d) \to \mathbb{L}^2(\mathcal{M}_d)$ is a linear (partial) differential operator, with order α_d defined appropriately.

Clearly, in order to apply Pen(u) to the candidate MPB, we must assume that $\xi_{k,d} \in \mathbb{W}^{\alpha_d,2}(\mathcal{M}_d)$. We penalize the total roughness for a set of $K > 1$ basis functions by summing the contribution from each ζ_k. We have the following result related to the representation of Pen:

Proposition 2.2. *There exists a symmetric positive semi-definite matrix $\boldsymbol{T}_d$, depending on L_d and ϕ_d, such that*

$$\sum_{k=1}^{K} \text{Pen}(\zeta_k) = \text{tr}(\tilde{\boldsymbol{C}}_d' \boldsymbol{T}_d \tilde{\boldsymbol{C}}_d), \qquad \tilde{\boldsymbol{C}}_d := \boldsymbol{D}_d \boldsymbol{V}_d' \boldsymbol{C}_d. \tag{14}$$

As a result, we have that penalties of the form of Definition 2.3 are quadratic in the transformed coordinate matrices $\tilde{\boldsymbol{C}}_d$ and therefore convex. Coupled with separability, this permits the derivation of efficient numerical algorithms to estimate the optimal MPB functions, which will be discussed in Section 2.2.3.

Penalization of the coefficient matrix $\boldsymbol{B}$ is incorporated using the penalty function denoted $l(\boldsymbol{B})$. We assume that $l(\boldsymbol{B})$ is convex, which is a requirement to guarantee the convergence of the algorithm in Section 2.2.3, but otherwise leave it's form unspecified. For example, common choices such as lasso and ridge penalties can be seamlessly integrated. Using the results from Theorem 2.1 and Proposition 2.2, the regularized augmentation of Equation (13) is a linear transformation of the solution to

$$\min_{\tilde{\boldsymbol{C}}_1,\ldots,\tilde{\boldsymbol{C}}_D,\boldsymbol{B}} \left\| \widehat{\mathcal{G}} - \sum_{k=1}^{K} \bigotimes_{d=1}^{D} \tilde{\boldsymbol{c}}_{d,k} \otimes \boldsymbol{b}_k \right\|_F^2 + \sum_{d=1}^{D} \lambda_d \text{tr}(\tilde{\boldsymbol{C}}_d' \boldsymbol{T}_d \tilde{\boldsymbol{C}}_d) + \lambda_{D+1} l(\boldsymbol{B}). \tag{15}$$

2.2.3 | Algorithm

The optimization problem (15) is non-convex 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

$$\tilde{\boldsymbol{C}}_d^{(r+1)} = \min_{\boldsymbol{X}} g(\tilde{\boldsymbol{C}}_1^{(r+1)}, \ldots, \tilde{\boldsymbol{C}}_{d-1}^{(r+1)}, \boldsymbol{X}, \tilde{\boldsymbol{C}}_{d+1}^{(r)}, \ldots, \tilde{\boldsymbol{C}}_D^{(r)}, \boldsymbol{B}^{(r)}), \tag{16}$$

for $d = 1, \ldots, D$ and likewise for $\boldsymbol{B}^{(r+1)}$, where g denotes the objective function from (15).

Using the properties of the d-mode matricization, we can write the conditional minimization problem defining the update of $\tilde{\boldsymbol{C}}_d$ as

$$\tilde{\boldsymbol{C}}_d^{(r+1)} = \min_{\tilde{\boldsymbol{C}}_d} \|\boldsymbol{G}_{(d)}' - \tilde{\boldsymbol{C}}_d \boldsymbol{W}_d^{(r)\prime}\|_F^2 + \lambda_d \text{tr}(\tilde{\boldsymbol{C}}_d' \boldsymbol{T}_d \tilde{\boldsymbol{C}}_d). \tag{17}$$

The update for $\boldsymbol{B}$ is given by

$$\boldsymbol{B}^{(r+1)} = \min_{\boldsymbol{B}} \|\boldsymbol{G}_{(D+1)} - \boldsymbol{W}_{D+1}^{(r)}\boldsymbol{B}'\|_F^2 + \lambda_{D+1} l(\boldsymbol{B}), \tag{18}$$

where $\boldsymbol{W}_d^{(r)} = (\bigodot_{j<d}^{D} \tilde{\boldsymbol{C}}_j^{(r+1)} \bigodot_{j>d}^{D} \tilde{\boldsymbol{C}}_j^{(r)}) \odot \boldsymbol{B}^{(r)}$ for $d = 1, \ldots, D$, $\boldsymbol{W}_{D+1}^{(r)} = \bigodot_{d=1}^{D} \tilde{\boldsymbol{C}}_d^{(r+1)}$ and $\boldsymbol{G}_{(d)}$ is the d-mode unfolding of tensor $\widehat{\mathcal{G}}$. Here $\odot$ is the Khatri–Rao product. From here on the superscript r denoting iteration is dropped for clarity.

In fact, the solution to the subproblem (17) is equivalent to the solution to

$$\tilde{\boldsymbol{C}}_d \boldsymbol{W}_d' \boldsymbol{W}_d + \lambda_d \boldsymbol{T}_d \tilde{\boldsymbol{C}}_d = \boldsymbol{W}_d' \boldsymbol{G}_{(d)}. \tag{19}$$

This equivalence can be verified by noting that (19) defines the gradient equations of (17) and that the solution is globally optimum due to convexity. Equation (19) is known as the Sylvester equation and has a unique solution under very mild conditions (specifically $\boldsymbol{W}_d'\boldsymbol{W}_d$ and $\lambda_d \boldsymbol{T}_d$ must have no common eigenvalues). Efficient algorithms for solving the Sylvester equation (Bartels and Stewart, 1972; Golub et al., 1979) are readily available in most common numerical computing languages.

Notice that by introducing the auxiliary variable $\boldsymbol{Z} = \boldsymbol{B}'$, the subproblem (18) can be written in separable form as

$$\begin{aligned} &\min_{\boldsymbol{B},\boldsymbol{Z}} \|\boldsymbol{G}_{(D+1)} - \boldsymbol{W}_{D+1}\boldsymbol{Z}\|_F^2 + \lambda_{D+1} l(\boldsymbol{B}) \\ &\text{subject to } \boldsymbol{B} - \boldsymbol{Z}' = \boldsymbol{0}. \end{aligned} \tag{20}$$

A numerical approximation to problems of the form (20) can be found using an alternating direction method of multipliers (ADMM) algorithm (Parikh and Boyd, 2014). The ADMM scheme consists of the iterates

$$\boldsymbol{B}_{\text{update}} \leftarrow \min_{\boldsymbol{B}} \left(\lambda_{D+1} l(\boldsymbol{B}) + \rho \|\boldsymbol{B} - \boldsymbol{Z}' + \boldsymbol{U}^*\|_F^2\right) \tag{21}$$

$$\boldsymbol{Z}_{\text{update}} \leftarrow \min_{\boldsymbol{Z}} \left(\|\boldsymbol{G}_{(D+1)} - \boldsymbol{W}_{D+1}\boldsymbol{Z}\|_F^2 + \rho \|\boldsymbol{B} - \boldsymbol{Z}' + \boldsymbol{U}^*\|_F^2\right) \tag{22}$$

$$\boldsymbol{U}^*_{\text{update}} \leftarrow \boldsymbol{U}^* + \boldsymbol{B} - \boldsymbol{Z}' \tag{23}$$

for some choice of $\rho > 0$, where $\boldsymbol{U}^*$ 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 (22) is a matrix ridge regression and has analytic solution given by

$$\boldsymbol{Z}_{\text{update}} = [\boldsymbol{W}_{D+1}'\boldsymbol{W}_{D+1} + \rho \boldsymbol{I}]^{-1}[\boldsymbol{W}_{D+1}'\boldsymbol{G}_{(D+1)} + \rho(\boldsymbol{B} + \boldsymbol{U}^*)']. \tag{24}$$

The update (21) 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 $\boldsymbol{Z}' - \boldsymbol{U}^*$.

Algorithm 1 provides pseudocode for the proposed block coordinate descent scheme. For the ADMM subproblem,

we adopt the stopping criteria proposed in Boyd et al. (2011) based on the primal and dual residuals at the r^{th} iteration, which have the form

$$r^{(r)}_{primal} = \|\boldsymbol{B}^{(r)} - \boldsymbol{Z}^{(r)\prime}\|_F, \qquad r^{(r)}_{dual} = \|\rho\left(\boldsymbol{Z}^{(r)} - \boldsymbol{Z}^{(r-1)}\right)\|_F. \tag{25}$$

To guarantee the convergence of Algorithm 1 to a stationary point of g and improve performance, an additional proximal regularization can be added to (16). Technical details are provided in Section S2, Supplementary Text.

We conclude this section with a few remarks on the practical implementation of Algorithm 1. Forming the matrix products $\boldsymbol{W}_d'\boldsymbol{W}_d$ and $\boldsymbol{W}_d'\boldsymbol{G}_{(d)}$ 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 $[\bigodot_i A_i]'[\bigodot_i A_i] = \bigcirc_i A_i'A_i$, where $\circ$ is the Hadamard product. Algorithms for efficient computation of the latter have been developed, see Phan et al. (2013). Following the suggestion of Huang et al. (2016), we found success setting $\rho = \boldsymbol{W}_{D+1}'\boldsymbol{W}_{D+1}/K$.

Algorithm 1 Algorithm to approximate the solution to (15)

Input $\widehat{\mathcal{G}}$, $\{\boldsymbol{T}_d\}$, $\{\lambda_d\}$
Output $\tilde{\boldsymbol{C}}_1, ..., \tilde{\boldsymbol{C}}_D$, $\boldsymbol{B}$
Initialize For $d = 1, ..., D$: $\tilde{\boldsymbol{C}}_d$; $\boldsymbol{U}^*$ as zero matrices; $\boldsymbol{B}$
while change in $\tilde{\boldsymbol{C}}_1, ..., \tilde{\boldsymbol{C}}_D$, $\boldsymbol{B}$ is non-negligible **do**
    for d = 1,...,D **do**
        Update $\tilde{\boldsymbol{C}}_d$ according to (17), by way of (19)
    while $r_{primal} > \text{tol}_{primal}$ or $r_{dual} > \text{tol}_{dual}$ **do**
        Update $\boldsymbol{B}$ according to (21)
        Update $\boldsymbol{Z}$ according to (24)
        Update $\boldsymbol{U}^*$ according to (23)
        Update r_{primal}, r_{dual} according to (25)

3 | THEORY

In the case of deterministic functions, convergence rates for optimal rank K approximations have been derived under various scenarios (Temlyakov, 2003; Barron et al., 2008). For random functions, the expected integrated squared error for representing realizations with the first K eigenfunctions is dictated by the decay rate of the spectrum of the covariance operator. In this section, we contribute to these results by analyzing the asymptotic approximation properties of the optimal rank-K MPB for representing realizations of a random function.

3.1 | Definitions and Assumptions

Without loss of generality, we assume the Lebesgue measure of $\mathcal{M}$ is 1. Note that under Assumption 1, we are guaranteed that the covariance function $C(\boldsymbol{x}, \boldsymbol{y}) := \mathbb{E}[U(\boldsymbol{x})U(\boldsymbol{y})]$ is continuous on $\mathcal{M} \times \mathcal{M}$. By Mercer's theorem, this covariance function has an eigen-decomposition $C(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{\infty} \rho_k \psi_k(\boldsymbol{x})\psi_k(\boldsymbol{y})$, where $\{\psi_k\}_{k=1}^{\infty}$ forms a complete orthonormal sequence of eigenfunctions in $\mathcal{H}$ and $\{\rho_k\}_{k=1}^{\infty}$ is a non-increasing sequence of real, non-negative eigenvalues.

Additionally, by the Karhunen-Loéve theorem, with probability one we have the decomposition $U(\boldsymbol{x}) = \sum_{k=1}^{\infty} Z_k \psi_k(\boldsymbol{x})$, where $Z_k = \langle U, \psi_k \rangle_{\mathcal{H}}$, which are mean zero random variables with $\mathbb{E}\left[Z_k Z_j\right] = \rho_k \delta_{kj}$.

Definition 3.1. Let the function $w_{\phi_d}(m_d)$ be the $\mathbb{L}^2(\mathcal{M}_d)$ convergence rate of the dth marginal basis system ϕ_d and $w_{\boldsymbol{\tau_m}}(\boldsymbol{m})$ be the $\mathbb{L}^2(\mathcal{M})$ convergence rate of the TPB system $\boldsymbol{\tau_m}$. That is, for any $f_d \in \mathcal{H}_d$, $f \in \mathcal{H}$

$$\left\| P_{\mathcal{H}^{\perp}_{m_d,d}}(f_d) \right\|_{\mathcal{H}_d} = O(w_{\phi_d}(m_d)), \qquad \left\| P_{\mathcal{H}^{\perp}_{\boldsymbol{m}}}(f) \right\|_{\mathcal{H}} = O(w_{\boldsymbol{\tau_m}}(\boldsymbol{m}))$$

where $P_{\mathcal{H}^{\perp}_{m_d,d}}$, $P_{\mathcal{H}^{\perp}_{\boldsymbol{m}}}$ are the projection operators onto $\mathcal{H}^{\perp}_{m_d,d}$ and $\mathcal{H}^{\perp}_{\boldsymbol{m}}$, the orthogonal complements of $\mathcal{H}_{m_d,d}$ in $\mathcal{H}_d$ and $\mathcal{H}_{\boldsymbol{m}}$ in $\mathcal{H}$, respectively.

Because ϕ_d is a complete basis of $\mathcal{H}_d$, $w_{\boldsymbol{\tau_m}}(\boldsymbol{m})$ is at least $o(1)$. The following assumption introduces conditions that are used to establish our convergence rate result.

Assumption 2.

$$\text{(i)} \sum_{k=K+1}^{\infty} \sqrt{\rho}_k = o(1) \quad \text{(ii)} \, \mathbb{E}\left[Z_k^4\right] < \infty, \ \forall k \quad \text{(iii)} \sum_{k=K+1}^{\infty} \mathbb{E}\left[Z_k^4\right]^{3/4} = o(1)$$

Assumption 2.i introduces a slightly stronger condition on the decay rate of the eigenvalues than the one that comes for free from Assumption 1, i.e. $\sum_{k=K+1}^{\infty} \rho_k = o(1)$. Assumption 2.ii is a standard moment condition. Assumption 2.iii is a technical condition which controls the fatness of the "high-frequency tail" of U. All three of these conditions are satisfied for many standard distributions and covariance kernels. The following two definitions describe objects required for the statement of our main results.

Definition 3.2. Let $\mathcal{A}_k$ be the the D-mode tensor with elements $\mathcal{A}_k(j_1, ..., j_D)$ defined by

$$P_{\mathcal{H}_{\boldsymbol{m}}}(\psi_k) = \sum_{j_1=1}^{m_1} \cdots \sum_{j_D=1}^{m_D} \mathcal{A}_k(j_1, ..., j_D) \phi_{1,j_1} \cdots \phi_{D,j_D},$$

where $P_{\mathcal{H}_{\boldsymbol{m}}}$ is the projection operator onto $\mathcal{H}_{\boldsymbol{m}}$.

Definition 3.3. Define the inner product space $(\bigotimes_{d=1}^{D} \mathbb{R}^{m_d}, \langle \cdot, \cdot \rangle_{\tilde{F}})$ where

$$\begin{aligned} \langle \mathcal{T}_1, \mathcal{T}_2 \rangle_{\tilde{F}} &= \langle \mathcal{T}_1, \mathcal{T}_2 \times_1 \boldsymbol{J}_{\phi_1} \cdots \times_D \boldsymbol{J}_{\phi_D} \rangle_F \equiv \langle \mathcal{T}_1 \times_1 \boldsymbol{J}_{\phi_1} \cdots \times_D \boldsymbol{J}_{\phi_D}, \mathcal{T}_2 \rangle_F \\ &= \sum_{i_1=1}^{m_1} \cdots \sum_{i_D=1}^{m_D} \sum_{j_1=1}^{m_1} \cdots \sum_{j_D=1}^{m_D} \mathcal{T}_1(i_1, ..., i_D) \mathcal{T}_2(j_1, ..., j_D) \prod_{d=1}^{D} \boldsymbol{J}_{\phi_d}(i_d, j_d) \end{aligned}$$

for tensors $\mathcal{T}_1, \mathcal{T}_2 \in \bigotimes_{d=1}^{D} \mathbb{R}^{m}$.

In order to ensure the existence and uniqueness of $\boldsymbol{\zeta}^*_m$ we must address the identifiability issue resulting from the permutation indeterminacy, i.e. the inherent ambiguity in the ordering of the basis functions. This issue can be resolved by imposing some ordering criteria on the parameters. In particular, we define the parameter space

$$\begin{aligned} \boldsymbol{\Theta}_{K,\boldsymbol{m}} := \{ &(\boldsymbol{C}_1, ..., \boldsymbol{C}_D) : \boldsymbol{c}'_{d,k} \boldsymbol{J}_{\phi_d} \boldsymbol{c}_{d,k} \leq 1 \text{ for } d = 1, ..., D; k = 1, ..., K \\ &\boldsymbol{C}_1(1,1) > ... > \boldsymbol{C}_D(1,K) \} \end{aligned}$$

which is a relaxation of $C_{K,\boldsymbol{m}}$ which resolves the permutation indeterminacy. We must also address a less superficial issue: the ill-posedness of best constrained rank approximations for $D > 2$ in general (de Silva and Lim, 2006). We invoke a sufficient but not necessary condition on K to resolve this issue (Sidiropoulos and Bro, 2000):

Assumption 3. Let $\mathcal{A}^{(K)}$ be the mode D+1 tensor obtained from stacking $\mathcal{A}_1, ..., \mathcal{A}_K$ for some finite integer K. Suppose it's rank is K^*. We assume that $K \geq (2K^* + D)/(D + 1)$.

3.2 | Asymptotic Results

Detailed proofs of the following results can be found in Section S1 of the Supplemental Materials. We establish the point-wise consistency and the convergence rate of $\check{\zeta}_{N,\boldsymbol{m}}$ in Theorems 3.1 and 3.2, respectively. Theorem 3.3 gives the generalization error of the K-oMPB and provides a quantification of the cost incurred compared to the true optimal rank K basis system.

Theorem 3.1 (Consistency). *Under Assumptions 1 and 3, we have the following (component-wise) convergence result:*

$$\check{\zeta}^*_{N,\boldsymbol{m}}(\boldsymbol{x}) \xrightarrow{P} \zeta^*_{\boldsymbol{m}}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathcal{M}.$$

Theorem 3.2 (Convergence Rate). *Under Assumptions 1, 2 and 3, we have*

$$\left\| \check{\zeta}^*_{N,\boldsymbol{m},k} - \zeta^*_{\boldsymbol{m},k} \right\|_{\mathcal{H}} = O_p(N^{-1/2}) \qquad \text{for each } k = 1, ..., K.$$

We now consider the expected generalization error of the K-oMPB. We compare this quantity to that obtained using the true optimal rank K basis system, i.e. the eigenfunctions. In doing so, we derive a form for quantifying the inefficiency cost of using the marginal product structure.

Theorem 3.3 (Generalization Error). *Let $\psi_K = (\psi_1, ..., \psi_K)'$ and denote P_{ψ_K} the projection operator onto span(ψ_K). Under Assumptions 1 and 3,*

$$\begin{aligned} \mathbb{E}\left\| U - P_{\zeta^*_{\boldsymbol{m}}}(U) \right\|^2_{\mathcal{H}} \quad &\leq \quad \text{Term}_1 + \text{Term}_2 + \text{Term}_3 \\ \text{Term}_1 &:= \mathbb{E}\left\| U - P_{\psi_K}(U) \right\|^2_{\mathcal{H}} = \sum_{k=K+1}^{\infty} \rho_k \\ \text{Term}_2 &:= \min_{C \in \boldsymbol{\Theta}_{K,\boldsymbol{m}}} \min_{\boldsymbol{B}} \sum_{l=1}^{K} \rho_l \left\| \mathcal{A}_l - \sum_{k=1}^{K} \boldsymbol{B}_{l,k} \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \right\|^2_{\tilde{F}} \\ \text{Term}_3 &:= O(w_{\tau_{\boldsymbol{m}}}(\boldsymbol{m})) \end{aligned} \tag{26}$$

Theorem 3.3 bounds the expected generalization error of the K-oMPB by the sum of three terms, the first term being the generalization error of the optimal rank K basis system. Hence the remaining two terms can be considered as a quantification of the cost incurred using the K-oMPB. Term$_3$ in Equation (26) comes from the irreducible contributions to the generalization error resulting from the finite truncation of the marginal ranks, i.e. the bias attached to using a finite basis expansion to represent an infinite dimensional object. Term$_2$, which is a tensor rank decomposition under a weighted augmentation of $\|\cdot\|_{\tilde{F}}$ (with weights determined by the variance of the corresponding components), quantifies the degree to which the leading modes of functional variation can be jointly represented using a basis system of MPFs. This term vanishes for large enough K, e.g. $K \geq K^*$, though this value will depend on the particular

tensor product space $\boldsymbol{\tau}$ as well as the marginal ranks $\boldsymbol{m}$. On the other hand, if the eigenfunctions themselves are rank-1 MPF, e.g. a sufficient condition being that the covariance function is separable, this finite term essentially vanishes for any K. This statement is made rigorous in the following corollary.

Corollary 3.4. *Assume that the eigenfunctions ψ_k are rank-1 MPF. Under the same set of assumptions, we have that*

$$\mathbb{E}\left\|U - P_{\hat{\zeta}^*_{\boldsymbol{m}}}(U)\right\|^2_{\mathcal{H}} = O\left(\prod_{d=1}^{D} w_{\phi_d}(m_d)\right) + \sum_{k=K+1}^{\infty} \rho_k \tag{27}$$

The separability assumption in Corollary 3.4 indicates that the eigenfunctions are in $\bigcup_{m\to\infty} \mathcal{V}_{m,K}$, so that the K-oMPB is equivalent to the leading K eigenfunctions. Hence, the generalization error only depends on the irreducible error from the finite truncation of the marginal basis and the tail-sum rate of the spectrum.

4 | SIMULATION STUDY

4.1 | Representing Random Marginal Product Functions

In this section, we compare three methods for constructing the functional representation of a random sample generated from a marginal product functional model: 1) a TPB system estimated by the sandwich smoother (Xiao et al., 2013), 2) the FCP-TPA algorithm (Allen, 2013), and 3) the K-oMPB estimated using Algorithm 1, referred to in this section as `MARGARITA` (MARGinal-product bAsis Representation wIth Tensor Analysis). A brief overview of the two competing methods is provided in Section S5, Supplementary Text.

The random function in our simulation is defined by the marginal product form;

$$U(\boldsymbol{x}) = \sum_{k=1}^{K^t} A^t_k \prod_{d=1}^{D} \left(\boldsymbol{c}^t_{d,k}\right)' \boldsymbol{\phi}^t_j(x_d).$$

Here $\boldsymbol{\phi}^t_j$ is the period-1 Fourier basis, $\boldsymbol{c}^t_{d,k}$ is the kth column vector of $\boldsymbol{C}^t_d$, the fixed marginal factor matrix such that each element is an $i.i.d.$ sample from $\mathcal{N}(0, 0.3^2)$; and $(A^t_1, ..., A^t_K)' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^t_A)$. In other words, U is a mean-zero Gaussian random field. The covariance matrix is constructed as $\boldsymbol{\Sigma}^t_A = \boldsymbol{ODO}'$, where $\boldsymbol{O}$ is a random $K^t \times K^t$ orthogonal matrix (sampled according to the Haar measure on $O(K^t)$), and $\boldsymbol{D}$ is a diagonal matrix with $\boldsymbol{D}_{kk} = \exp(-0.7k)$ for $k = 1, \ldots, K^t$, i.e. an exponential decay model of the spectrum. We took the function domain to be the unit cube $\mathcal{M} = [0, 1]^3$. We fixed the true marginal basis dimensions to be $m^t_d = 11$ 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 vs low SNR; obtained by taking of σ^2 to be 0.5 or 10, small vs. large domain sample size; $n_d = 30$ or 50 for all d, respectively, and small vs. 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 (11). The performance of the fitting methods are assessed by computing the mean integrated squared error (MISE) for each replication. That is, for each replication r, an estimate of the MISE is

$$\text{MISE}^{(r)} = \sum_{i=1}^{N^{(r)}} \int_{[0,1]^3} \left[U^{(r)}_i(\boldsymbol{x}) - \widehat{U}^{(r)}_i(\boldsymbol{x})\right]^2 \mathrm{d}\boldsymbol{x},$$

where $\widehat{U}^{(r)}_i$ is an estimate of $U^{(r)}_i$ from the rth simulated dataset. Denote the Monte Carlo average of the MISE as

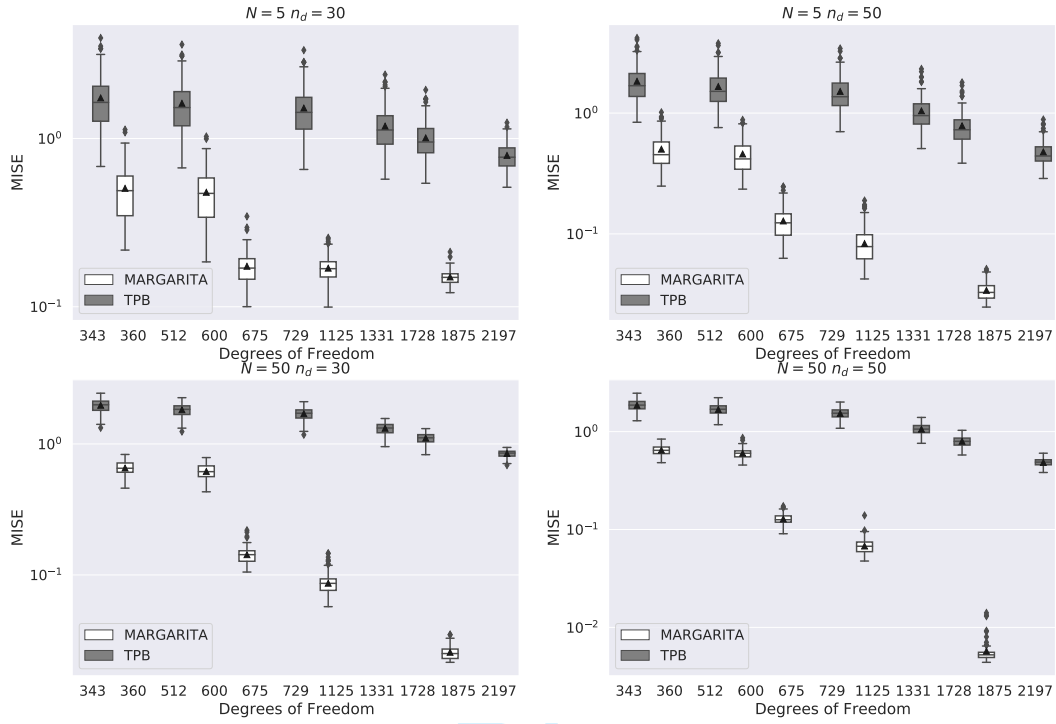


FIGURE 1 Comparison of the fit performance measured by MISE between the TPB estimated by the sandwhich smoother (gray) and `MARGARITA` (white) as a function of the total number of degrees of freedom. For each panel $K_t = 20$ and $\sigma^2 = 10$, but similar patterns emerge for the $K_t = 10$ and $\sigma^2 = 0.5$ cases as well. The Y-axis is plotted on log-scale for clarity.

$\text{moMISE} = 100^{-1} \sum_{r=1}^{100} \text{MISE}^{(r)}$.

We believe a fair comparison between the TPB and `MARGARITA` should be based on enforcing (roughly) equivalently sized parameter spaces, i.e. total number of degrees of freedom (DF), which are $\prod_{d=1}^{D} m_d$ and $K_{\text{fit}} \sum_{d=1}^{D} m_d$ for the TPB and `MARGARITA`, respectively. For the former we use a tensor product of marginal cubic B-splines with marginal rank set equal to the smallest integer m_d such that $m_d^3 \geq K_{\text{fit}} \sum_{d=1}^{D} m_d$. For the latter, we considered $K_{\text{fit}} = 8, 15,$ or 25 and use a marginal cubic B-spline basis of rank 15 or 25. The second order derivative was used to define the marginal roughness penalties and a ridge penalty was used for regularization on the coefficients. For both FCP-TPA and `MARGARITA`, the penalty parameters were chosen through a grid search. For the TPB, the smoothing parameters were selected by minimizing the GCV criterion from Xiao et al. (2013), implemented in R package `hero` (French, 2020). Simulations were performed using R/4.0.2 and Python/3.7.7 on a Linux machine equipped with a 2.4 GHz Intel Xeon CPU E5-2695 and 24GB of RAM.

Figure 1 shows a comparison of the fit performance of the sandwhich smoother and `MARGARITA` for various simulation settings. Overall, the `MARGARITA` fits had substantially lower moMISE compared to the TPB fits with comparable DF. Further discussion of these results as well as a tabular display of the moMISE for each simulation setting and model parameterization, are provided in Section S5 and Table S1 in the Supplementary Text. For the remainder of this section, we focus on comparing FCP-TPA and `MARGARITA`.

Figure 2 shows boxplots of the MISE for FCP-TPA and `MARGARITA` for each combination of marginal and global

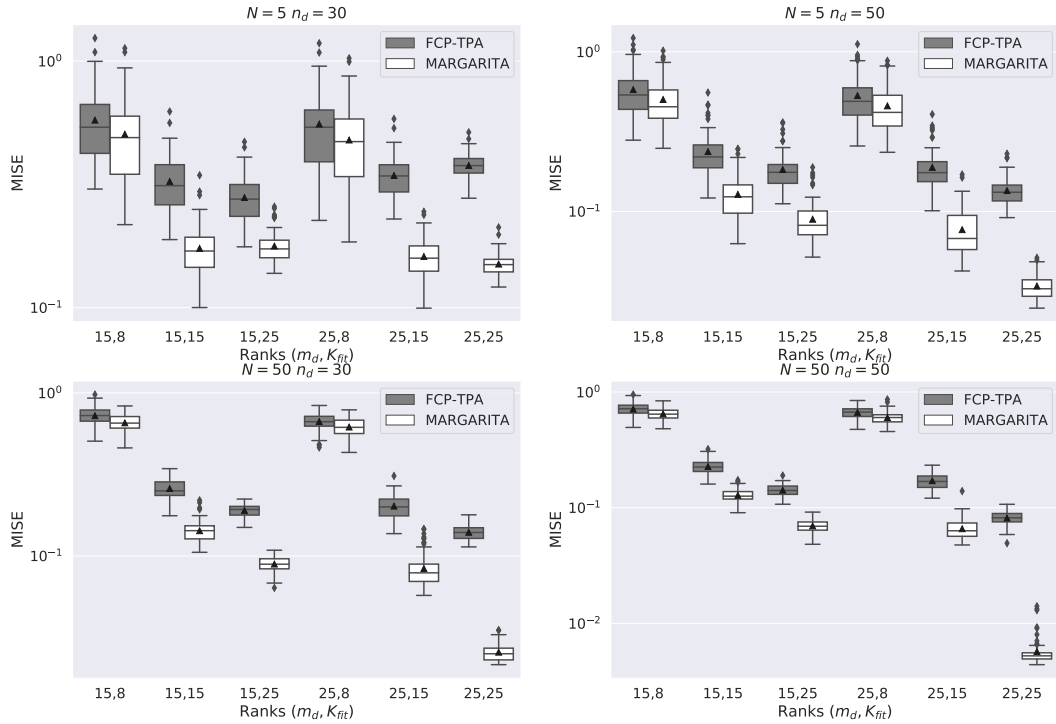


FIGURE 2 MISE of marginal product fits resulting from both FCP-TPA (gray) and `MARGARITA` (white). moMISE is denoted by a triangle. The Y-axis is plotted on log-scale for clarity.

ranks, m_d and K_{fit} respectively, for several combinations of sample size N and marginal domain size n_d. For all panels, $K_t = 20$ and $\sigma^2 = 10$. In all cases, `MARGARITA` results in fits with lower moMISE than FCP-TPA. The separation in performance is more apparent in models with larger rank. An exhaustive tabular comparison of the moMISE for FCP-TPA and `MARGARITA` for all simulation settings considered is provided in Table S2 of Section S5, Supplementary Text.

We also compared the computational time of FCP-TPA and `MARGARITA` for each of the simulated datasets. For the small sample size, small sample domain case, the computational speed is comparable between the two algorithms. As both N and n_d increase, `MARGARITA` begins to comparatively fare better, culminating in significantly faster performance in the $N = 50$, $n_d = 50$ case. These results are shown in Figure S1, Supplementary Text. This trend is expected, since increasing n_d does not increase the dimension of the optimization problem (15) at the heart of Algorithm 1. On the other hand, 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. This highlights the practical importance of our methods ability to control the dimension of the optimization problem.

4.2 | Generalization Performance

The previous section evaluates the in-sample fitting performance of our method. We are also interested in the generalization performance of `MARGARITA`, that is, how efficiently does the K-oMPB estimated from a sample of size N_{train} represent new realizations from the same distribution? The results of Section 3 indicate that we can expect

near optimal generalization performance, with an inefficiency that vanishes for increasing K. We compare our a method to the marginal product FPCA procedure proposed in Chen et al. (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, and so in this study we let the functional domain be $\mathcal{M} = [0, 1]^2$. The eigenfunctions determining a non-stationary, non-separable anisotropic covariance function $C(\mathbf{x}, \mathbf{y})$ over $\mathcal{M}$ are defined as follows. Denote the eigen-decomposition of the pairwise $\mathbb{L}^2$ inner product matrix of the tensor product basis system $\boldsymbol{J}_{\phi_1 \otimes \phi_2} = \boldsymbol{P}\boldsymbol{\Gamma}\boldsymbol{P}'$. The collection of $m_1 m_2$ orthonormal eigenfunctions are defined according to $\psi = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{P}'\text{vec}(\phi_1 \otimes \phi_2)$. The eigenvalue corresponding to the kth eigenfunction is given by an exponential decay model $\rho_k = \exp(-0.7k)$. Realizations of the random function $U_i \sim U$ are simulated using a Gaussian process assumption and then evaluated on an equispaced 100×100 grid on $\mathcal{M}$.

We are interested in evaluating the generalization error of both `MARGARITA` and `MargFPCA` for a variety of training sample sizes N_{train} and ranks K. For simplicity, ϕ_1 and ϕ_2 are used as the marginal basis for fitting and are taken to be equispaced cubic b-splines with $m_1 = 10$ and $m_2 = 8$. The second order derivative operator is used to define the marginal roughness penalty and the coefficients are penalized with a ridge penalty. For each combination of N_{train} and K, both the K-oMPB and the marginal eigenfunctions, estimated by `MARGARITA` and `MargFPCA` respectively, are used to construct the representations for each of 50 realizations from an independent test set using the least squares principle. The MISE is then approximated using numerical integration over a dense grid. Each experimental set-up is repeated for 25 replications.

Figure 3 displays average MISE as a function K for both `MARGARITA` (red) and `MargFPCA` (blue). The dotted and solid lines correspond to $N_{train} = 10$ and $N_{train} = 100$, respectively. The performance of the first K true eigenfunctions is included for comparison. The MISE curves are uniformly shifted down for the larger N_{train}, as expected. 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. Section S5.4 of the Supplemental Material provides the results for a variety N_{train}, which result in similar conclusions.

If it is desirable to obtain estimates of the eigenfunctions, `MARGARITA` can be utilized for this purpose as part of a fast two-stage procedure. This approach first computes the K-oMPB and then performs FPCA using the represented data, i.e. representing the eigenfunctions as linear combinations of the K-oMPB. See Section S3 in the Supplemental Material for more details on this procedure. For each simulated dataset, we estimate the first three eigenfunctions using the proposed two-stage procedure with K-oMPB of rank 60. In order to avoid the sign ambiguity associated with the eigenfunctions, we evaluated these estimates using the absolute inner product: $\text{AIP} = |\langle \psi_k, \hat{\psi}_k \rangle_{\mathcal{H}}|$. Table 1 displays the results for several N_{train}. We see that the AIP approaches 1 for all three eigenfunctions as N_{train} increases. Section S5.4 of the Supplemental Material provides plots of the estimated and true eigenfunctions for the large sample case.

5 | REAL DATA ANALYSIS

The white matter (WM) of the human brain consists of large collections of myelinated nueral fibers that permit fast communication between disparate regions of the brain. Diffusion magnetic resonance imaging (dMRI) is a non-invasive imaging technique which uses spatially localized measurements of the diffusion of water molecules to allow researches to probe the WM microstructure. At each 3-dimensional voxel in the brain, the diffusion image can be used to compute scalar summaries of local diffusion, e.g. fractional anistropy (FA) or mean diffusivity. The resulting data

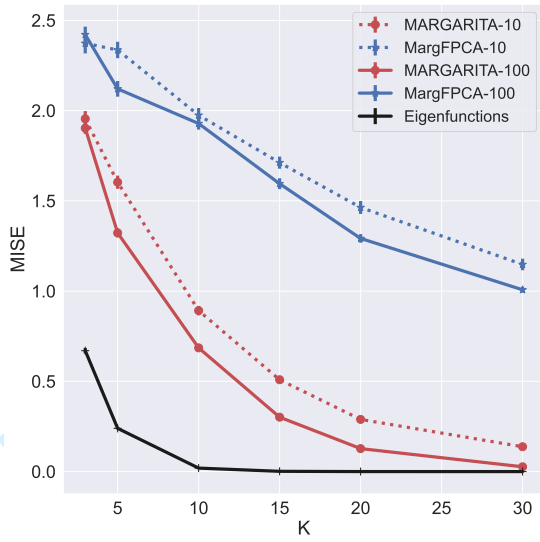


FIGURE 3 Comparison of the generalization performance as a function of K for both `MARGARITA` (red) and `MargFPCA` (blue), for $N_{train} = 10$ (dotted) $N_{train} = 100$ (solid). The true optimal rank K basis system, i.e. the eigenfunctions, are included for comparison (solid black line).

	AIP for several N_{train}		
	$N_{train} = 50$	$N_{train} = 100$	$N_{train} = 200$
ψ_1	0.9147 (0.0335)	0.9708 (0.0065)	0.9877 (0.0021)
ψ_2	0.8397 (0.0413)	0.9543 (0.0078)	0.9779 (0.0026)
ψ_3	0.8614 (0.0320)	0.9391 (0.0138)	0.9776 (0.0032)

TABLE 1 Average absolute inner product (AIP) between the true and two-stage estimates for the first three eigenfunctions. Standard errors are given in parenthesis.

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; Asselin et al., 2020). Typically, voxel-based analysis are performed for group-wise analysis of FA (Smith et al., 2006). This results in an enormous multiple testing problem and, as a result, many voxel-based analysis of FA in TBI studies are not able to establish significant group differences (Khong et al., 2016). On the other hand, 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 Equation (11). In this analysis, we focus on a functional approach to 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.

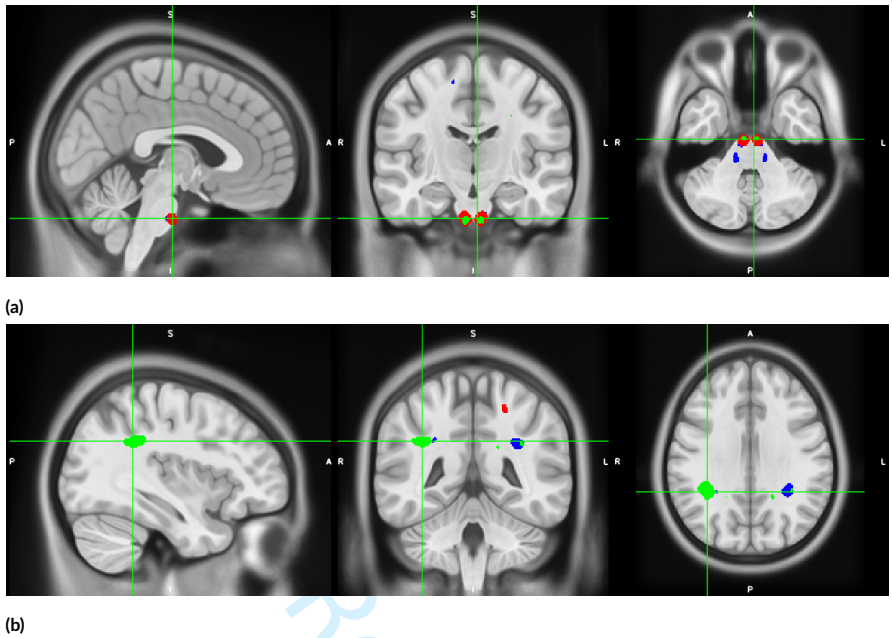
After preprocessing, the final voxel grid is of size $115 \times 140 \times 120$. Point-wise estimates of the mean function at

each voxel are obtained using the sample mean tensor, which is then used to center the data. Equispaced cubic b-splines of ranks $m_1 = 100$, $m_2 = 120$, $m_3 = 100$ are used as marginal basis systems. Marginal roughness is penalized by the second order derivative. A K-oMPB of rank $K = 330$ is fit to the mean centered data tensor using `MARGARITA`. The penalty parameters are selected using a grid search while the marginal and global ranks are selected by sequentially training larger models until more than 85% of the variance in the raw data is captured. FPCA is performed on the represented data using the approach outlined in Section S3 of the Supplementary Material. The first 45 eigenfunctions, denoted collectively as ψ, 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 ψ. The resulting classification performance is evaluated using leave-one-out cross validation (CV). Note that in order to more faithfully simulate the clinical setting, the left-out subject is not included in the estimation of the K-oMPB. In order to obtain the coefficient vector for the left-out subject, a least squares estimator is used to project the raw-FA data into the span of ψ. Finally, the data driven regions of interest (ROIs) are defined as spatial volumes where the values of the 3 most informative eigenfunctions (as determined by the classifier) are "extreme", i.e. outside the 0.5% and 99.5% quantiles.

The cross validated accuracy, precision and recall are approximately 0.88, 0.85, and 0.92, respectively, indicating substantial discriminatory power of the learned basis functions. Figure 4 displays two cross sections of the template space. The blue, red and green areas indicate data driven regions of interest (ROIs). The red and green ROIs in Figure 4(a) are within areas of the middle cerebellar peduncle (MCP), a structure composed of multiple fibers mostly involved in motor processing (Morales and Tomsick, 2015). The green and blue ROIs in Figure 4(b) are within areas along the superior longitudinal fasciculus (SLF), a fiber bundle that is involved in higher-order motor and language processing (Petrides and Pandya, 2002). Wang et al. (2016) found increased FA in the MCP is associated with increased cognitive impairment. Xiong 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. In addition, Brandstack et al. (2013) found changes in long WM tracts in the brain using tractography on subjects with TBI. Therefore, our findings of changes in the MCP and SLF are consistent with the hypothesized mechanism and previous findings in TBI.

6 | DISCUSSION AND FUTURE WORK

Our work introduces a methodological framework and accompanying estimation algorithm for constructing an optimal 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. Theoretical properties of interest are also established. 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



(a)

(b)

FIGURE 4 Data-driven ROIs created from thresholding the 0.5% and 99.5% quantiles of the three most informative eigenfunctions, blue, red, and green, respectively. ROIs in plot (a) are within areas of the middle cerebellar peduncle and in plot (b) are in areas along the superior longitudinal fasciculus, both of which are known to be affected in TBI.

of diffusion MRI, we show that our method can facilitate the prediction of disease status and identify biologically meaningful ROIs.

One of the key ways the proposed method differs from alternative multidimensional tensor smoothing methods is the explicit estimation of a continuous set of basis functions. This continuous representation of the data permits computationally efficient operations such as inner products and partial derivatives, from which fast procedures for more complex operations can be built, e.g. FPCA, functional regression, partial differential equation modeling. In the big data era, it is increasingly the case that large collections of publicly available historically relevant data can be acquired for many applications of interest, e.g. Human Connectome Project for brain MRI (Glasser et al., 2016). It is our vision that the proposed methodology can be applied to such historical databases and the learned basis systems ported to subsequent analysis of data hypothesized to come from the same population. This pipeline will dramatically reduce both the computation requirements and the dimensionalities of the resulting models. Note that one set of raw diffusion data in this study contained roughly 1.9 million voxels. Using our methods, it can be mapped down to a dimension of 45, an approximately 40,000 times reduction, while still retaining enough signal to be able to predict disease status with high accuracy.

A major advantage of the proposed method is its computational efficiency, as discussed in Section 4 and further illustrated in Figure S1 of the Supplementary Text. The coordinate transformation at the heart of Theorem 2.1 allows the user to control the dimension of the optimization problem. Simulation results recorded in Section S5 of the Supplementary Material indicate large savings in computational time compared to a tensor decomposition of the raw data, for moderately large grids. Another benefit of the proposed method is its modular structure. The marginal basis system, differential regularization and coefficient penalty function are all relatively generic, which allows for a variety

of different specifications that can be customized to the problem at hand. We implemented our methods in a Python package named **eMFDA**, which is available at `https://github.com/Will-Consagra/eMFDA`. Our software is built for easy interface with the `scikit-fda` package: `https://github.com/GAA-UAM/scikit-fda`, and hence can take advantage of the large class of differential penalties and basis systems implemented therein.

We now discuss a few practical considerations of our methodology that may arise. First, in Equation (11), we assume that the functional samples are observed on a common grid. In practice, this assumption may be overly restrictive: data may not be available at a subset of grid points for some samples, i.e. missing data. We could augment the loss function in Equation (15) to accommodate some missing values. This would clearly change the form of the updates (17) and (18), but algorithms have been proposed for solving similar problems (Huang et al., 2016). On the computational side, our method requires the specification of $D + 1$ penalty parameters. In both simulation and real data experiments, we have found that performing CV over a grid of values centered at $\|\boldsymbol{W}_d'\boldsymbol{W}_d\|_F / \|\boldsymbol{T}_d\|_F$ is typically sufficient for selecting a good penalty strength. Finally, we assume that the functional samples are mean zero. In the real data application, the point-wise mean of the observed tensors was used to center the data. Another option is to estimate an explicit functional form. Simple adjustments can be made to our methodology to estimate a separate marginal product representation for the deterministic mean function, see Section S5.5 of the Supplemental Text for details.

This work can be extended in several interesting directions. As discussed, the penalty parameter selection is currently being performed through a grid search and therefore a principled derivation of a generalized cross validation criteria for our method is of interest. Up to this point, we have considered the ϕ_d's as fixed user-supplied parameters. In some cases, problem-specific prior information can be used to guide the selection of the marginal basis systems, e.g. Fourier basis if $\mathcal{M}_d = \mathbb{S}^1$. Additionally, if the number of marginal grid points n_d is not too large, using a locally supported basis, e.g. splines, with $m_d = n_d$ and then promoting smoothness through selecting λ_d can be a good strategy. That said, we are often interested in the case when n_d is large and or $\mathcal{M}_d$ is not an interval. In these more complicated situations, it is likely the case that optimizing over the marginal basis systems would result in substantially better performance of the resulting K-oMPB. We discuss one possible extension of the proposed method to permit data-adaptive marginal basis systems using random projections in Section S4, Supplementary Text, but more work is needed. From a theoretical perspective, the results established in Section 3 can be broadened. Our theory is derived under the fully observed case and does not take into account the effect of discretization over a grid. Understanding the asymptotic behavior in this finite resolution regime is of interest.

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)*, 220–223.

Asselin, P. D., Gu, Y., Merchant-Borna, K., Abar, B., Wright, D. W., Qiu, X. and Bazarian, J. J. (2020) Spatial regression analysis of mr diffusion reveals subject-specific white matter changes associated with repetitive head impacts in contact sports. *Scientific reports*, **10**, 1–12.

Aston, J. A. D., Pigoli, D. and Tavakoli, S. (2017) Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, **45**, 1431–1461. URL: `http://www.jstor.org/stable/26362871`.

Barron, A. R., Cohen, A., Dahmen, W. and DeVore, R. A. (2008) Approximation and learning by greedy algorithms. *The Annals of Statistics*, **36**, 64 – 94. URL: `https://doi.org/10.1214/009053607000000631`.

Bartels, R. H. and Stewart, G. W. (1972) Solution of the matrix equation AX + XB = C. *Commun. ACM*, **15**, 820–826. URL: `https://doi.org/10.1145/361573.361582`.

Beylkin, G., Garcke, J. and Mohlenkamp, M. J. (2009) Multivariate regression and machine learning with sums of separable functions. *SIAM Journal on Scientific Computing*, **31**, 1840–1857. URL: https://doi.org/10.1137/070710524.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3**, 1–122. URL: https://doi.org/10.1561/2200000016.

Brandstack, N., Kurki, T. and Tenovuo, O. (2013) Quantitative diffusion-tensor tractography of long association tracts in patients with traumatic brain injury without associated findings at routine mr imaging. *Radiology*, **267**, 231–239.

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 (Statistical Methodology)*, **79**, 177–196. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12160.

Chen, L.-H. and Jiang, C.-R. (2017) Multi-dimensional functional principal component analysis. *Statistics and Computing*, **27**, 1181–1192. URL: https://doi.org/10.1007/s11222-016-9679-5.

Chevreuil, M., Lebrun, R., Nouy, A. and Rai, P. (2015) A least-squares method for sparse low rank approximation of multivariate functions. *SIAM/ASA Journal on Uncertainty Quantification*, **3**, 897–921. URL: https://doi.org/10.1137/13091899X.

French, J. (2020) *hero: Spatio-Temporal (Hero) Sandwich Smoother*. URL: https://CRAN.R-project.org/package=hero. R package version 0.4.7.

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L., Auerbach, E. J., Behrens, T. E., Coalson, T. S., Harms, M. P., Jenkinson, M., Moeller, S., Robinson, E. C., Sotiropoulos, S. N., Xu, J., Yacoub, E., Ugurbil, K. and Van Essen, D. C. (2016) The Human Connectome Project's neuroimaging approach. *Nature Neuroscience*, **19**, 1175–1187.

Golub, G., Nash, S. and Van Loan, C. (1979) A hessenberg-schur method for the problem AX + XB = C. *IEEE Transactions on Automatic Control*, **24**, 909–913.

Gorodetsky, A., Karaman, S. and Marzouk, Y. (2019) A continuous analogue of the tensor-train decomposition. *Computer Methods in Applied Mechanics and Engineering*, **347**, 59–84. URL: https://www.sciencedirect.com/science/article/pii/S0045782518306133.

Grasedyck, L., Kressner, D. and Tobler, C. (2013) A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, **36**, 53–78. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gamm.201310004.

Hillar, C. J. and Lim, L.-H. (2013) Most tensor problems are np-hard. *J. ACM*, **60**. URL: https://doi.org/10.1145/2512329.

Hsing, T. and Eubank, R. (2015) *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

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. URL: https://doi.org/10.1198/jasa.2009.tm08024.

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.

Kargas, N. and Sidiropoulos, N. D. (2021) Supervised learning and canonical decomposition of multivariate functions. *IEEE Transactions on Signal Processing*, **69**, 1097–1107.

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. URL: https://www.frontiersin.org/article/10.3389/fneur.2016.00156.

Kolda, T. G. and Bader, B. W. (2009) Tensor decompositions and applications. *SIAM review*, **51**, 455–500.

Kraus, M. F., Susmaras, T., Caughlin, B. P., Walker, C. J., Sweeney, J. A. and Little, D. M. (2007) White matter integrity and cognition in chronic traumatic brain injury: a diffusion tensor imaging study. *Brain*, **130**, 2508–2519. URL: https://doi.org/10.1093/brain/awm216.

Li, Y., Huang, C. and Härdle, W. K. (2019) Spatial functional principal component analysis with applications to brain image data. *Journal of Multivariate Analysis*, **170**, 263 – 274. URL: http://www.sciencedirect.com/science/article/pii/S0047259X1730742X.

Liang, D., Huang, H., Guan, Y. and Yao, F. (2021) Test of weak separability for spatially stationary functional field. *Journal of the American Statistical Association*, **0**, 1–43. URL: https://doi.org/10.1080/01621459.2021.2002156.

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. URL: https://doi.org/10.1093/biomet/asy048.

Masak, T. and Panaretos, V. M. (2019) Spatiotemporal covariance estimation by shifted partial tracing.

Masak, T., Sarkar, S. and Panaretos, V. M. (2020) Principal separable component analysis via the partial inner product.

Morales, H. and Tomsick, T. (2015) Middle cerebellar peduncles: Magnetic resonance imaging and pathophysiologic correlate. *World journal of radiology*, **7**, 438.

Nouy, A. (2017) Low-rank tensor methods for model order reduction. *Handbook of Uncertainty Quantification*, 857–882. URL: http://dx.doi.org/10.1007/978-3-319-12385-1_21.

Parikh, N. and Boyd, S. (2014) Proximal algorithms. *Found. Trends Optim.*, **1**, 127–239. URL: https://doi.org/10.1561/2400000003.

Petrides, M. and Pandya, D. N. (2002) Association pathways of the prefrontal cortex and functional observations. *Principles of frontal lobe function*, **1**, 31–50.

Phan, A., Tichavský, P. and Cichocki, A. (2013) Fast alternating LS algorithms for high order CANDECOMP/PARAFAC tensor factorizations. *IEEE Transactions on Signal Processing*, **61**, 4834–4846.

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.

Ramsay, J. and Silverman, B. W. (2005) *Functional Data Analysis*. Springer. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa239.

Ruppert, D., Wand, M. and Carroll, R. (2006) *Semiparametric Regression*. Cambridge University Press.

Sidiropoulos, N. D. and Bro, R. (2000) On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, **14**, 229–239.

de Silva, V. and Lim, L.-H. (2006) Tensor rank and the ill-posedness of the best low-rank approximation problem.

Silverman, B. W. (1996) Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, **24**, 1–24. URL: https://doi.org/10.1214/aos/1033066196.

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. URL: https://www.sciencedirect.com/science/article/pii/S1053811906001388.

Stone, C. J. (1980) Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, **8**, 1348–1360. URL: http://www.jstor.org/stable/2240947.

Suzuki, T., Kanagawa, H., Kobayashi, H., Shimizu, N. and Tagami, Y. (2016) Minimax optimal alternating minimization for kernel nonparametric tensor learning. In *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/file/b4568df26077653eeadf29596708c94b-Paper.pdf.

Temlyakov (2003) Nonlinear methods of approximation. *Foundations of Computational Mathematics*, **3**, 33–107. URL: https://doi.org/10.1007/s102080010029.

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*, **0**, 1–14. URL: https://doi.org/10.1080/01621459.2020.1820344.

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.

Wasserman, L. (2010) *All of Nonparametric Statistics*. Springer Publishing Company, Incorporated, 1st edn.

Xiao, L., Li, Y. and Ruppert, D. (2013) Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 577–599. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12007.

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.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590. URL: https://doi.org/10.1198/016214504000001745.

Zapata, J., Oh, S.-Y. and Petersen, A. (2020) Partial separability and functional graphical models for multivariate gaussian processes.

Zhang, J.-T. and Chen, J. (2007) Statistical inferences for functional data. *The Annals of Statistics*, **35**, 1052 – 1079. URL: https://doi.org/10.1214/009053606000001505.