ORIGINAL ARTICLE

Supplementary Text for: Efficient Multidimensional Functional Data Analysis Using Marginal Product Basis Systems

Remark: Equations defined in this supplementary text are all labeled by "Equation (S<n>)" convention, so that they can be distinguished from those defined in the main text, which do not have prefix "S".

S1 | THEORY AND PROOFS

S1.1 | Proof of Proposition 2.2

Proof. Denote the matrix $\boldsymbol{R}_d(i,j) = \int_{\mathcal{M}_d} L_d(\phi_{d,i})L_d(\phi_{d,j})$. Then we have

$$\begin{aligned}\int_{\mathcal{M}_d} L_d^2(\xi_{k,d}) &= \boldsymbol{c}_{d,k}'\boldsymbol{R}_d\boldsymbol{c}_{d,k} \\ &= \tilde{\boldsymbol{c}}_{d,k}'\boldsymbol{D}_d^{-1}\boldsymbol{V}_d'\boldsymbol{R}_d\boldsymbol{V}_d\boldsymbol{D}_d^{-1}\tilde{\boldsymbol{c}}_{d,k} \\ &\equiv \tilde{\boldsymbol{c}}_{d,k}'\boldsymbol{T}_d\tilde{\boldsymbol{c}}_{d,k}.\end{aligned}$$

The result follows by noting that $\boldsymbol{T}_d$ is a symmetric product of positive semidefinite matrices and therefore is symmetric and positive semidefinite. □

S1.2 | Consistency

In this section we establish the point-wise consistency of $\check{\zeta}^*_{\boldsymbol{m},N}$. Throughout this section, let $0 < R < \infty$ be a generic constant, that is perhaps different depending on context. For notational convenience, we establish the following definition.

Definition S1.1. Let the function $h(K)$ be convergence rate of the tail-sum of the eigenvalues of the covariance operator associated with $C(\boldsymbol{x},\boldsymbol{y})$, that is

$$\sum_{k=K+1}^{\infty} \rho_k = O(h(K))$$

Proposition S1. *The coefficients of the projection $P_{\boldsymbol{\zeta_m}}(U) := \sum_{k=1}^{K} b_k \xi_k$ are given by*

$$\boldsymbol{b}_K := (b_1, ..., b_K)' = \sum_{l=1}^{\infty} Z_l \boldsymbol{b}_l$$

where

$$\boldsymbol{b}_l = \begin{pmatrix} \prod_{d=1}^{D} \boldsymbol{c}'_{d,1} \boldsymbol{J}_{\phi_d} \boldsymbol{c}_{d,1} & \prod_{d=1}^{D} \boldsymbol{c}'_{d,1} \boldsymbol{J}_{\phi_d} \boldsymbol{c}_{d,2} & \cdots & \prod_{d=1}^{D} \boldsymbol{c}'_{d,1} \boldsymbol{J}_{\phi_d} \boldsymbol{c}_{d,K} \\ \prod_{d=1}^{D} \boldsymbol{c}'_{d,2} \boldsymbol{J}_{\phi_d} \boldsymbol{c}_{d,1} & \prod_{d=1}^{D} \boldsymbol{c}'_{d,2} \boldsymbol{J}_{\phi_d} \boldsymbol{c}_{d,2} & \cdots & \\ \vdots & & \ddots & \\ \prod_{d=1}^{D} \boldsymbol{c}'_{d,K} \boldsymbol{J}_{\phi_d} \boldsymbol{c}_{d,1} & & \cdots & \prod_{d=1}^{D} \boldsymbol{c}'_{d,K} \boldsymbol{J}_{\phi_d} \boldsymbol{c}_{d,K} \end{pmatrix}^{-1} \begin{bmatrix} \langle \mathcal{A}_l, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,1} \rangle_{\tilde{F}} \\ \langle \mathcal{A}_l, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,2} \rangle_{\tilde{F}} \\ \vdots \\ \langle \mathcal{A}_l, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,K} \rangle_{\tilde{F}} \end{bmatrix} \tag{S.1}$$

Proof. This follows from the definition of the $\mathbb{L}^2$ projection operator. □

Proposition S2. *$\|\boldsymbol{b}_l\|_2^2 < R$ for all l.*

Proof. Note that the norm of the matrix inverse in (S.1) is bounded due to the linear independence of the ξ_k's. Additionally, each $|\langle \mathcal{A}_l, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \rangle_{\tilde{F}}| \leq 1$ from Cauchy–Schwarz. The desired result follows immediately. □

Proposition S3. *We have that (i) $\mathbb{E}\left[| \langle U, P_{\boldsymbol{\zeta_m}}(U) \rangle_{\mathcal{H}} | \right] < R$ and (ii) $\mathbb{E}\left[\|P_{\boldsymbol{\zeta_m}}(U)\|_{\mathcal{H}}^2 \right] < R$ for any $\boldsymbol{\zeta_m} \in \mathcal{V}_{K,\boldsymbol{m}}$.*

Proof. For (i), notice that

$$\begin{aligned} \langle U, P_{\boldsymbol{\zeta_m}}(U) \rangle_{\mathcal{H}} &= \left\langle \sum_{q=l}^{\infty} Z_q \psi_q, \sum_{k=1}^{K} b_k \xi_k \right\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{q=l}^{\infty} Z_q \psi_q, \sum_{l=1}^{\infty} Z_l \boldsymbol{b}'_l \boldsymbol{\zeta_m} \right\rangle_{\mathcal{H}} \\ &= \sum_{q=1}^{\infty} \sum_{l=1}^{\infty} Z_q Z_l \left\langle \mathcal{A}_q, \sum_{k=1}^{K} b_{l,k} \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \right\rangle_{\tilde{F}} . \end{aligned}$$

Taking the expectation and re-arragning terms we have,

$$\mathbb{E}\left[| \langle U, P_{\boldsymbol{\zeta_m}}(U) \rangle_{\mathcal{H}} | \right] \leq \sum_{l=1}^{\infty} \sum_{k=1}^{K} \rho_l |b_{l,k}| |\langle \mathcal{A}_l, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \rangle_{\tilde{F}}| < R$$

where the bound follows from applying proposition S2, Cauchy–Schwarz and that $\sum_{l=1}^{\infty} \rho_l < \infty$
Likewise, for (ii)

$$\begin{aligned} \mathbb{E}\left[\|P_{\boldsymbol{\zeta_m}}(U)\|_{\mathcal{H}}^2 \right] &= \sum_{l=1}^{\infty} \rho_l \boldsymbol{b}'_l \boldsymbol{J}_{\boldsymbol{\zeta_m}} \boldsymbol{b}_l \\ &\leq \sum_{l=1}^{\infty} \rho_l \, \|\boldsymbol{b}_l\|_2^2 \, \|\boldsymbol{J}_{\boldsymbol{\zeta_m}}\|_F^2 < R \end{aligned}$$

where $\boldsymbol{J}_{\boldsymbol{\zeta_m}} = \int_{\mathcal{M}} \boldsymbol{\zeta_m} \boldsymbol{\zeta}'_{\boldsymbol{m}}$. □

Lemma S4. *Let $\mathcal{A}^{(K)}$ be the mode $D+1$ tensor obtained from stacking $\mathcal{A}_1, \ldots, \mathcal{A}_K$. Define the inner product space $\left(\bigotimes_{d=1}^{D} \mathbb{R}^{m_d} \otimes \mathbb{R}^K, \langle \cdot, \cdot \rangle_{\tilde{F},C}\right)$, where*

$$\langle \mathcal{T}_1, \mathcal{T}_2 \rangle_{\tilde{F},C} = \sum_{k=1}^{K} \rho_k \langle \mathcal{T}_1(:, \ldots, :, k), \mathcal{T}_2(:, \ldots, :, k) \rangle_{\tilde{F}}.$$

Then the expected generalization error of the $\boldsymbol{\zeta}_m$ can be written as

$$\mathbb{E} \left\| U - P_{\boldsymbol{\zeta_m}}(U) \right\|_{\mathcal{H}}^2 = \min_{\boldsymbol{B}} \quad \left\| \mathcal{A}^{(K)} - \sum_{k=1}^{K} \left[\bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \right] \otimes \boldsymbol{B}_{:,k} \right\|_{\tilde{F},C}^2 + O(w_{\boldsymbol{\tau_m}}(\boldsymbol{m})) + O(h(K)) \tag{S.2}$$

where $\boldsymbol{B}_{:,k}$ is the k'th column of $\boldsymbol{B} \in \mathbb{R}^{K \times K}$.

Proof.

$$\begin{aligned} \mathbb{E} \left\| U - P_{\boldsymbol{\zeta_m}}(U) \right\|_{\mathcal{H}}^2 &= \mathbb{E} \left\| P_{\mathcal{H}_{\boldsymbol{m}}}(U) - P_{\boldsymbol{\zeta_m}}(U) + P_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}(U) \right\|_{\mathcal{H}}^2 \\ &= \mathbb{E} \left\| P_{\mathcal{H}_{\boldsymbol{m}}}(U) - P_{\boldsymbol{\zeta_m}}(U) \right\|_{\mathcal{H}_{\boldsymbol{m}}} + \\ &+ \mathbb{E} \left\langle (P_{\mathcal{H}_{\boldsymbol{m}}}(U) - P_{\boldsymbol{\zeta_m}}(U)), P_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}(U) \right\rangle_{\mathcal{H}} \\ &+ \mathbb{E} \left\| P_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}(U) \right\|_{\mathcal{H}_{\boldsymbol{m}}^{\perp}} \\ &:= \text{Term}_1 + \text{Term}_2 + \text{Term}_3. \end{aligned}$$

Term$_3$ is independent of $\boldsymbol{\zeta_m}$ and represents the expected irreducible error due to the finite dimensional truncation of the marginal basis systems. We have

$$\begin{aligned} \text{Term}_3 = \mathbb{E} \left\| P_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}(U) \right\|_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}^2 &= \mathbb{E} \left\| \sum_{k=1}^{\infty} Z_k P_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}(\psi_k) \right\|_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}^2 \\ &= \sum_{k=1}^{\infty} \mathbb{E} \left[Z_k^2 \right] \cdot \left\| P_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}(\psi_k) \right\|_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}^2 \\ &= O(w_{\boldsymbol{\tau_m}}(\boldsymbol{m})), \end{aligned} \tag{S.3}$$

where the second line follows from the Z_k being uncorrelated and the third line follows since $\sum_{k=1}^{\infty} \mathbb{E}\left[Z_k^2\right] = \sum_{k=1}^{\infty} \rho_k < \infty$. Since span$(\boldsymbol{\zeta_m}) \subset \mathcal{H}_{\boldsymbol{m}}$, it is easy to see that Term$_2 = 0$ and thus we need only to deal with Term$_1$.

The mapping $\iota : \mathcal{H}_{\boldsymbol{m}} \mapsto \bigotimes_{d=1}^{D} \mathbb{R}^{m_d}$ defined by $\iota(u)_{j_1, \ldots, j_D} = a_{j_1, \ldots, j_D}$ is an isometry between inner product spaces $(\mathcal{H}_{\boldsymbol{m}}, \langle \cdot, \cdot \rangle_{\mathcal{H}_{\boldsymbol{m}}})$ and $(\bigotimes_{d=1}^{D} \mathbb{R}^{m_d}, \langle \cdot, \cdot \rangle_{\tilde{F}})$, where $a_{j_1, \ldots, j_D}$ is the coefficient of u associated with basis element $\prod_{d=1}^{D} \phi_{d, j_d}$. Recall that any $u \in$ span$(\boldsymbol{\zeta_m})$ has the representation

$$u(\boldsymbol{x}) = \sum_{k=1}^{K} b_k \prod_{d=1}^{D} \sum_{j=1}^{m_d} c_{k,d,j} \phi_{d,j}(x_d)$$

and hence, under ι, is identified with the tensor rank-K tensor $\sum_{k=1}^{K} b_k \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k}$, where $\boldsymbol{c}_{d,k}$ are the m_d-vectors of coefficients for the kth marginal function in the dth dimension. It follows that

$$\begin{aligned}
\left\|P_{\mathcal{H}_m}(U) - P_{\zeta_m}(U)\right\|^2_{\mathcal{H}_m} &= \left\|\sum_{l=1}^{\infty} Z_l \mathcal{A}_l - \sum_{k=1}^{K} b_k \bigotimes_{d=1}^{D} \mathbf{c}_{d,k}\right\|^2_{\tilde{F}} \\
&= \mathbb{E}\left\|\sum_{l=1}^{\infty} Z_l \mathcal{A}_l - \sum_{k=1}^{K}\sum_{j=1}^{\infty} Z_l b_{j,k} \bigotimes_{d=1}^{D} \mathbf{c}_{d,k}\right\|^2_{\tilde{F}} \\
&= \mathbb{E}\left\|\sum_{l=1}^{\infty} Z_l \mathcal{A}_l - \sum_{k=1}^{K} Z_l b_{l,k} \bigotimes_{d=1}^{D} \mathbf{c}_{d,k}\right\|^2_{\tilde{F}} \\
&= \sum_{l=1}^{\infty} \mathbb{E}\left[Z_l^2\right] \left\|\mathcal{A}_l - \sum_{k=1}^{K} b_{l,k} \bigotimes_{d=1}^{D} \mathbf{c}_{d,k}\right\|^2_{\tilde{F}} \\
&\quad + \sum_{j \neq r} \mathbb{E}\left[Z_j Z_r\right] \left\langle \mathcal{A}_j - \sum_{k=1}^{K} b_{j,k} \bigotimes_{d=1}^{D} \mathbf{c}_{d,k}, \mathcal{A}_r - \sum_{k=1}^{K} b_{r,k} \bigotimes_{d=1}^{D} \mathbf{c}_{d,k} \right\rangle_{\tilde{F}} \\
&= \sum_{l=1}^{\infty} \rho_l \left\|\mathcal{A}_l - \sum_{k=1}^{K} b_{l,k} \bigotimes_{d=1}^{D} \mathbf{c}_{d,k}\right\|^2_{\tilde{F}} \\
&= \sum_{l=1}^{K} \rho_l \left\|\mathcal{A}_l - \sum_{k=1}^{K} b_{l,k} \bigotimes_{d=1}^{D} \mathbf{c}_{d,k}\right\|^2_{\tilde{F}} + O(h(K)) \\
&= \min_{\mathbf{B}} \sum_{k=1}^{K} \rho_l \left\|\mathcal{A}_l - \sum_{k=1}^{K} \mathbf{B}_{l,k} \bigotimes_{d=1}^{D} \mathbf{c}_{d,k}\right\|^2_{\tilde{F}} + O(h(K)) \\
&= \min_{\mathbf{B}} \left\|\mathcal{A}^{(K)} - \sum_{k=1}^{K} \left[\bigotimes_{d=1}^{D} \mathbf{c}_{d,k}\right] \otimes \mathbf{B}_{:,k}\right\|^2_{\tilde{F},C} + O(h(K)).
\end{aligned}$$

□

Lemma S5. *Let*

$$L_N(\boldsymbol{C}) := N^{-1} \sum_{i}^{N} \|U_i - P_{\boldsymbol{C}}(U_i)\|^2_{\mathcal{H}}; \quad L(\boldsymbol{C}) := \mathbb{E}\, \|U - P_{\boldsymbol{C}}(U)\|^2_{\mathcal{H}}.$$

where $P_{\boldsymbol{C}}$ is the reparameterization of the projection operator P_{ζ_m} for ζ_m defined by $\boldsymbol{C} = (\boldsymbol{C}_1, ..., \boldsymbol{C}_D) \in \boldsymbol{\Theta}_{K,m}$. Define $\check{\boldsymbol{C}}_N, \boldsymbol{C}^* \in \boldsymbol{\Theta}_{K,m}$ to be the minimizers of $L_N(\boldsymbol{C})$ and $L(\boldsymbol{C})$, respectively. Then

$$\check{\boldsymbol{C}}_N \xrightarrow{P} \boldsymbol{C}^*$$

Proof. The strong law of large numbers ensures $L_N(\boldsymbol{C}) \to L(\boldsymbol{C})$ for every $\boldsymbol{C}$, almost surely. By Theorem 5.7 of Vaart (1998), the desired convergence holds if the following conditions are met:

1. *Uniform Convergence*:

$$\sup_{\boldsymbol{C} \in \boldsymbol{\Theta}_{K,m}} |L_N(\boldsymbol{C}) - L(\boldsymbol{C})| \xrightarrow{P} 0$$

2. *Uniqueness*: For any $\epsilon > 0$

$$\sup_{\boldsymbol{C}:\text{dist}(\boldsymbol{C},\boldsymbol{C}^*)\geq\epsilon} L(\boldsymbol{C}) > L(\boldsymbol{C}^*)$$

3. *Near Minimum*:

$$L_N(\check{\boldsymbol{C}}_N) \leq L_N(\boldsymbol{C}^*) + o_P(1)$$

Condition 1: This can be verified by using Glivenko-Cantelli theory. Denote $l_{\boldsymbol{C}}(U) = \|U - P_{\boldsymbol{C}}(U)\|^2_{\mathcal{H}}$, i.e. $L(\boldsymbol{C}) = \mathbb{E}\left[l_{\boldsymbol{C}}(U)\right]$. Denote the set of functions

$$\Gamma = \{l_{\boldsymbol{C}} : \boldsymbol{C} \in \boldsymbol{\Theta}_{K,\boldsymbol{m}}\}.$$

The uniform convergence requirement is equivalent to Γ being Glivenko-Cantelli. We can express each element of the function set as

$$l_{\boldsymbol{C}}(u) = \|u\|^2_{\mathcal{H}} - 2\langle u, P_{\boldsymbol{C}}(u)\rangle_{\mathcal{H}} + \langle P_{\boldsymbol{C}}(u), P_{\boldsymbol{C}}(u)\rangle_{\mathcal{H}}$$

Here, we work with the equivalent formulation of

$$\begin{aligned} l_{\boldsymbol{C}}(u) &= \|u\|^2_{\mathcal{H}} - 2\langle u, P_{\boldsymbol{C}}(u)\rangle_{\mathcal{H}} + \langle P_{\boldsymbol{C}}(u), P_{\boldsymbol{C}}(u)\rangle_{\mathcal{H}} \\ &= \sum_{l=1}^{\infty} Z_q^2 - 2\sum_{k=1}^{K} b_k \langle \sum_{q=1}^{\infty} Z_q \mathcal{A}_q, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \rangle_{\tilde{F}} + \sum_{k=1}^{K}\sum_{p=1}^{K} b_k b_p \langle \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k}, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,jp} \rangle_{\tilde{F}} \end{aligned} \tag{S.4}$$

Recalling the definition of $\langle , \rangle_{\tilde{F}}$ and we have

$$\begin{aligned} \left\langle \sum_{q=1}^{\infty} Z_q \mathcal{A}_q, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \right\rangle_{\tilde{F}} &= \sum_{i_1=1}^{m_1} \cdots \sum_{i_D=1}^{m_d} \sum_{j_1=1}^{m_1} \cdots \sum_{j_D=1}^{m_D} \left(\sum_{q=1}^{\infty} Z_q \mathcal{A}_q(i_1, ..., i_D) \right) \prod_{d=1}^{D} \boldsymbol{J}_{\phi_d}(i_d, j_d) \prod_{d=1}^{D} \boldsymbol{c}_{d,k,j_d} \\ \left\langle \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k}, \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,p} \right\rangle_{\tilde{F}} &= \sum_{i_1=1}^{m_1} \cdots \sum_{i_D=1}^{m_d} \sum_{j_1=1}^{m_1} \cdots \sum_{j_D=1}^{m_D} \prod_{d=1}^{D} \boldsymbol{J}_{\phi_d}(i_d, j_d) \prod_{d=1}^{D} \boldsymbol{c}_{d,k,i_d} \boldsymbol{c}_{d,p,j_d} \end{aligned}$$

which are polynomials of order D and $2D$ in $\boldsymbol{C}$, respectively. From the definition given in proposition S1, we can see that each b_k is also a finite degree polynomial in $\boldsymbol{C}$. As a result, $l_{\boldsymbol{C}}(u)$ is isomorphic to a polynomial with finitely many terms. From the boundedness of the sum of the second moments of the Z_k's along with proposition S3, it follows that $\mathbb{E}\left[l_{\boldsymbol{C}}\right] < \infty$. Therefore, the Γ is VC-class, which follows from Lemma 2.6.16 of van der Vaart and Wellner (1996), and hence Glivenko-Cantelli.

Condition 2: This condition indicates $\boldsymbol{C}^*$ is a well separated minimum of L. Using Lemma S.2, we have that

$$\min_{\zeta_{\boldsymbol{m}} \in \mathcal{V}_{K,\boldsymbol{m}}} \mathbb{E}\left\|U - P_{\zeta_{\boldsymbol{m}}}(U)\right\|^2_{\mathcal{H}} = \min_{\boldsymbol{C} \in \boldsymbol{\Theta}_{K,\boldsymbol{m}}} \min_{\boldsymbol{B}} \left\| \mathcal{A}^{(K)} - \sum_{k=1}^{K} \left[\bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \right] \otimes \boldsymbol{B}_{:,k} \right\|^2_{\tilde{F},C} + O(w_{\mathcal{T}_{\boldsymbol{m}}}(\boldsymbol{m})) + O(h(K))$$

and therefore the minimizer of L is given by the rank K decomposition of the tensor $\mathcal{A}^{(K)}$ under the $\| \cdot \|_{\tilde{F},C}$ norm.

Under Assumption 4, this minimizer is unique in $\boldsymbol{\Theta}_{K,\boldsymbol{m}}$. Coupled with the compactness of $\boldsymbol{\Theta}_{K,\boldsymbol{m}}$ and continuity of L, the desired condition follows.

Condition 3: This follows trivially, as

$$L_N(\breve{\boldsymbol{C}}_N) = \min_{\boldsymbol{C}\in\boldsymbol{\Theta}_{K,\boldsymbol{m}}} N^{-1}\sum_i^N \|U_i - P_{\boldsymbol{C}}(U_i)\|_{\mathcal{H}}^2 \leq N^{-1}\sum_i^N \|U_i - P_{\boldsymbol{C}^*}(U_i)\|_{\mathcal{H}}^2 = L_N(\boldsymbol{C}^*)$$

□

Proof of Theorem 3.1

Proof. Note that $\zeta_{\boldsymbol{m}}(\boldsymbol{x})$ is a continuous functions of $\boldsymbol{C}$ for all $\boldsymbol{x} \in \mathcal{M}$. The desired result follows directly from the convergence established in Lemma S5 and the continuous mapping theorem. □

S1.3 | Convergence Rate

Lemma S6 (Lipschitz Map). *There exists functional $F(u)$ such that $\mathbb{E}\left[F^2\right] < \infty$ and*

$$|l_{\boldsymbol{C}^{(1)}}(u) - l_{\boldsymbol{C}^{(2)}}(u)| \leq F(u) dist(\boldsymbol{C}^{(1)}, \boldsymbol{C}^{(2)})$$

for any $\boldsymbol{C}^{(1)}, \boldsymbol{C}^{(2)} \in \Theta_{K,\boldsymbol{m}}$, where $dist(\boldsymbol{C}_1, \boldsymbol{C}_2) = \sum_{k=1}^K \left\|\bigotimes_{d=1}^D \mathbf{c}_{d,k}^{(1)} - \bigotimes_{d=1}^D \mathbf{c}_{d,k}^{(2)}\right\|_{\tilde{F}}$.

Proof. Notice that

$$\begin{aligned}
|l_{\boldsymbol{C}^{(1)}}(u) - l_{\boldsymbol{C}^{(2)}}(u)| &= |-2\langle u, P_{\boldsymbol{C}^{(1)}}(u) - P_{\boldsymbol{C}^{(2)}}(u)\rangle_{\mathcal{H}} + \|P_{\boldsymbol{C}^{(1)}}(u)\|_{\mathcal{H}}^2 - \|P_{\boldsymbol{C}^{(2)}}(u)\|_{\mathcal{H}}^2| \\
&\leq 2|\langle u, P_{\boldsymbol{C}^{(1)}}(u) - P_{\boldsymbol{C}^{(2)}}(u)\rangle_{\mathcal{H}}| + |\|(P_{\boldsymbol{C}^{(1)}}(u)\|_{\mathcal{H}} - \|P_{\boldsymbol{C}^{(2)}}(u)\|_{\mathcal{H}})(P_{\boldsymbol{C}^{(1)}}(u)\|_{\mathcal{H}} + \|P_{\boldsymbol{C}^{(2)}}(u)\|_{\mathcal{H}})| \\
&\leq 2\|u\|_{\mathcal{H}}\|P_{\boldsymbol{C}^{(1)}}(u) - P_{\boldsymbol{C}^{(2)}}(u)\|_{\mathcal{H}} + \|P_{\boldsymbol{C}^{(1)}}(u) - P_{\boldsymbol{C}^{(2)}}(u)\|_{\mathcal{H}}(\|P_{\boldsymbol{C}^{(1)}}(u)\|_{\mathcal{H}} + \|P_{\boldsymbol{C}^{(2)}}(u)\|_{\mathcal{H}}) \\
&\leq 4\|u\|_{\mathcal{H}}\|P_{\boldsymbol{C}^{(1)}}(u) - P_{\boldsymbol{C}^{(2)}}(u)\|_{\mathcal{H}}.
\end{aligned}$$

Additionally, we have that

$$\begin{aligned}
\|P_{\boldsymbol{C}_1}(u) - P_{\boldsymbol{C}_2}(u)\|_{\mathcal{H}} &= \|\sum_{l=1}^{\infty} Z_l \sum_{k=1}^K (b_{l,k}^{(1)} \bigotimes_{d=1}^D \mathbf{c}_{d,k}^{(1)} - b_{l,k}^{(2)} \bigotimes_{d=1}^D \mathbf{c}_{d,k}^{(2)})\|_{\tilde{F}} \\
&\leq \sum_{l=1}^{\infty} |Z_l| \sum_{k=1}^K \|(b_{l,k}^{(1)} \bigotimes_{d=1}^D \mathbf{c}_{d,k}^{(1)} - b_{l,k}^{(2)} \bigotimes_{d=1}^D \mathbf{c}_{d,k}^{(2)})\|_{\tilde{F}} \\
&\leq R \sum_{l=1}^{\infty} |Z_l| \sum_{k=1}^K \|\bigotimes_{d=1}^D \boldsymbol{c}_{d,k}^{(1)} - \bigotimes_{d=1}^D \boldsymbol{c}_{d,k}^{(2)}\|_{\tilde{F}}.
\end{aligned}$$

Define $F(u) := R\|u\|_{\mathcal{H}} \sum_{l=1}^{\infty} |Z_l|$ for generic constant $0 < R < \infty$. We have that

$$
\begin{aligned}
\mathbb{E}\left[F^2\right] &= R\mathbb{E}\left[\left(\sum_{l=1}^{\infty} Z_l^2\right)\left(\sum_{l=1}^{\infty} |Z_l|\right)^2\right] \\
&= R\mathbb{E}\left[\sum_{l=1}^{\infty}\sum_{j=1}^{\infty}\sum_{q=1}^{\infty} Z_l^2 |Z_j||Z_q|\right] \\
&= R\sum_{l=1}^{\infty}\sum_{j=1}^{\infty}\sum_{q=1}^{\infty} \mathbb{E}\left[Z_l^2\right]\mathbb{E}\left[|Z_j|\right]\mathbb{E}\left[|Z_q|\right]\mathbb{I}\{l \neq j, l \neq q, q \neq j\}+ \\
&\qquad \mathbb{E}\left[Z_l^2\right]\mathbb{E}\left[Z_q^2\right]\mathbb{I}\{l \neq j, j = q\}+ \\
&\qquad \mathbb{E}\left[|Z_l|^3\right]\mathbb{E}\left[|Z_j|\right]\mathbb{I}\{l \neq j, l = q\}+ \\
&\qquad \mathbb{E}\left[|Z_l|^3\right]\mathbb{E}\left[|Z_q|\right]\mathbb{I}\{l = j, l \neq q\}+ \\
&\qquad \mathbb{E}\left[Z_l^4\right]\mathbb{I}\{l = j = q\}.
\end{aligned}
$$

Therefore, $\sum_{l=1}^{\infty} \mathbb{E}\left[|Z_l|^r\right] < \infty$ for $r = 1, 2, 3, 4 \implies \mathbb{E}\left[F^2\right] < \infty$. We have $\mathbb{E}\left[|Z_l|\right] \leq \sqrt{\rho_l}$ by Jensen's inequality, and $\mathbb{E}\left[|Z_l|^3\right] \leq \mathbb{E}\left[Z_l^4\right]^{3/4}$ by Holder's inequality. These bounds along with Assumptions 1 and 3 ensure each of these series are convergent and the desired result follows. □

Proof of Theorem 3.2

Proof. Using lemmas S6 and S5 along with corollary 5.53 of Vaart (1998) and recalling the definition of $\|\cdot\|_{\tilde{F}}$, we have

$$
\left\|\check{\boldsymbol{\zeta}}^*_{N,\boldsymbol{m},k} - \boldsymbol{\zeta}^*_{\boldsymbol{m},k}\right\|_{\mathcal{H}} = \left\|\bigotimes_{d=1}^{D} \check{\boldsymbol{c}}_{N,d,k} - \bigotimes_{d=1}^{D} \boldsymbol{c}^*_{d,k}\right\|_{\tilde{F}} = O_p(N^{-1/2}) \text{ for each } k = 1, \ldots, K
$$

as desired. □

S1.4 | Generalization Error

Proof of Theorem 3.3

Proof. We have that

$$
\begin{aligned}
\mathbb{E}\left\|U - P_{\boldsymbol{\zeta}^*_{\boldsymbol{m}}}(U)\right\|_{\mathcal{H}}^2 &= \mathbb{E}\left\|U - P_{\boldsymbol{\zeta}^*_{\boldsymbol{m}}}(U) + P_{\boldsymbol{\psi}_K}(U) - P_{\boldsymbol{\psi}_K}(U)\right\|_{\mathcal{H}}^2 \\
&\leq \mathbb{E}\left\|U - P_{\boldsymbol{\psi}_K}(U)\right\|_{\mathcal{H}}^2 + \mathbb{E}\left\|P_{\boldsymbol{\psi}_K}(U) - P_{\boldsymbol{\zeta}^*_{\boldsymbol{m}}}(U)\right\|_{\mathcal{H}}^2
\end{aligned}
\tag{S.5}
$$

Considering now the second term in the sum on line two, observe that

$$
\begin{aligned}
\mathbb{E}\left\|P_{\boldsymbol{\psi}_K}(U) - P_{\boldsymbol{\zeta}^*_{\boldsymbol{m}}}(U)\right\|_{\mathcal{H}}^2 &= \mathbb{E}\left\|P_{\mathcal{H}_m}\left(P_{\boldsymbol{\psi}_K}(U) - P_{\boldsymbol{\zeta}^*_{\boldsymbol{m}}}(U)\right) + P_{\mathcal{H}_m^{\perp}}\left(P_{\boldsymbol{\psi}_K}(U) - P_{\boldsymbol{\zeta}^*_{\boldsymbol{m}}}(U)\right)\right\|_{\mathcal{H}}^2 \\
&= \mathbb{E}\left\|P_{\mathcal{H}_m}\left(P_{\boldsymbol{\psi}_K}(U)\right) - P_{\boldsymbol{\zeta}^*_{\boldsymbol{m}}}(U)\right\|_{\mathcal{H}_m}^2 + \mathbb{E}\left\|P_{\mathcal{H}_m^{\perp}}\left(P_{\boldsymbol{\psi}_K}(U)\right)\right\|_{\mathcal{H}_m^{\perp}}^2 \\
&= \text{Term}_1 + \text{Term}_2
\end{aligned}
$$

Clearly, $\text{Term}_2 = O(w_{\mathcal{T}_{\boldsymbol{m}}}(\boldsymbol{m}))$. In regard to Term_1, using the same logic as in the proof of Lemma S4, we have that

$$\begin{aligned}\text{Term}_1 &= \min_{\boldsymbol{B}} \sum_{k=1}^{K} \rho_l \left\| \mathcal{A}_l - \sum_{k=1}^{K} \boldsymbol{B}_{l,k} \bigotimes_{d=1}^{D} \boldsymbol{c}^*_{d,k} \right\|_{\tilde{F}}^2 \\ &= \min_{\boldsymbol{C} \in \boldsymbol{\Theta}_{K,\boldsymbol{m}}} \min_{\boldsymbol{B}} \sum_{l=1}^{K} \rho_l \left\| \mathcal{A}_l - \sum_{k=1}^{K} \boldsymbol{B}_{l,k} \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \right\|_{\tilde{F}}^2\end{aligned}$$

where the $O(h(K))$ term is avoided due to the finite truncation of ψ_K. The desired results follows from plugging the derived forms of Term_1 and Term_2 into Equation S.5. □

Proof of Corollary 3.4

Proof. Under the separability assumption, for all k we have $\psi_k(\mathbf{x}) = \prod_{d=1}^{D} \psi_{k,d}(x_d)$ and hence

$$P_{\mathcal{H}_{m,d}}(\psi_{k,d}) = \prod_{d=1}^{D} \sum_{j=1}^{m_d} a_{d,k,j} \phi_{d,j}$$

for some set of coefficients $\{a_{d,k,j}\}$. Define the vector $\boldsymbol{a}_{d,k} := (a_{d,k,1}, ..., a_{d,k,m_d})' \in \mathbb{R}^{m_d}$. Notice that

$$\mathcal{A}_k = \bigotimes_{d=1}^{D} \boldsymbol{a}_{d,k}$$

and therefore

$$\min_{\boldsymbol{C} \in \boldsymbol{\Theta}_{K,\boldsymbol{m}}} \min_{\boldsymbol{B}} \sum_{k=1}^{K} \rho_l \left\| \mathcal{A}_l - \sum_{k=1}^{K} \boldsymbol{B}_{l,k} \bigotimes_{d=1}^{D} \boldsymbol{c}_{d,k} \right\|_{\tilde{F}}^2 \leq \sum_{k=1}^{K} \rho_l \left\| \mathcal{A}_l - \sum_{k=1}^{K} \boldsymbol{I}_{l,k} \bigotimes_{d=1}^{D} \boldsymbol{a}_{d,k} \right\|_{\tilde{F}}^2 = 0.$$

The proof is completed by noting that, under the separable assumption, Term_2 from the proof to Theorem 3.3 becomes

$$\begin{aligned}\mathbb{E} \left\| P_{\mathcal{H}_{\boldsymbol{m}}^{\perp}} \left(P_{\psi_K}(U) \right) \right\|_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}^2 &= \mathbb{E} \left\| \sum_{k=1}^{K} Z_k P_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}(\psi_k) \right\|_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}^2 \\ &= \mathbb{E} \left\| \sum_{k=1}^{K} Z_k \prod_{d=1}^{D} P_{\mathcal{H}_{m_d,d}^{\perp}}(\psi_{k,d}) \right\|_{\mathcal{H}_{\boldsymbol{m}}^{\perp}}^2 \\ &= \sum_{k=1}^{K} \rho_k \prod_{d=1}^{D} \left\| P_{\mathcal{H}_{m_d,d}^{\perp}}(\psi_{k,d}) \right\|_{\mathcal{H}_{m_d}^{\perp}}^2 \\ &= O\left(\prod_{d=1}^{D} w_{\phi_d}(m_d) \right)\end{aligned}$$

□

S1.5 | Proof of Theorem 2.1

Proof.

$$\begin{aligned}\sum_{i=1}^{N}\left\|\mathcal{Y}_i-\sum_{k=1}^{K}\boldsymbol{B}_{ik}\bigotimes_{d=1}^{D}\boldsymbol{\Phi}_d\boldsymbol{c}_{d,k}\right\|_F^2 &= \left\|\mathcal{Y}-\sum_{k=1}^{K}\bigotimes_{d=1}^{D}\boldsymbol{\Phi}_d\boldsymbol{c}_{d,k}\otimes\boldsymbol{b}_k\right\|_F^2\\ &= \left\|\mathcal{Y}-\Big[\sum_{k=1}^{K}\bigotimes_{d=1}^{D}\boldsymbol{D}_d\boldsymbol{V}_d'\boldsymbol{c}_{d,k}\otimes\boldsymbol{b}_k\Big]\times_1\boldsymbol{U}_1\times_2\boldsymbol{U}_2\cdots\times_D\boldsymbol{U}_D\right\|_F^2\\ &= \left\|\widehat{\mathcal{G}}-\Big[\sum_{k=1}^{K}\bigotimes_{d=1}^{D}\boldsymbol{D}_d\boldsymbol{V}_d'\boldsymbol{c}_{d,k}\otimes\boldsymbol{b}_k\Big]\right\|_F^2.\end{aligned}$$

Here the first equality is from properties of the Frobenius norm, the second comes from properties of d-mode multiplication, and the third from invariance of the Frobenius norm to orthogonal transformation. Therefore, solving Equation (13) is equivalent to solving

$$\min_{\boldsymbol{B},\boldsymbol{C}}\left\|\widehat{\mathcal{G}}-\sum_{k=1}^{K}\bigotimes_{d=1}^{D}\boldsymbol{D}_d\boldsymbol{V}_d'\boldsymbol{C}_d\otimes\boldsymbol{b}_k\right\|_F^2. \tag{S.6}$$

Using the mapping $\tilde{\boldsymbol{C}}_d=\boldsymbol{D}_d\boldsymbol{V}_d'\boldsymbol{C}_d$, Equation (S.6) can be reparameterized as

$$\min_{\boldsymbol{B},\tilde{\boldsymbol{C}}}\left\|\widehat{\mathcal{G}}-\sum_{k=1}^{K}\bigotimes_{d=1}^{D}\tilde{\boldsymbol{C}}_d\otimes\boldsymbol{b}_k\right\|_F^2, \tag{S.7}$$

which is solved by the rank-K CPD of $\widehat{\mathcal{G}}$. Comparing Equations (S.6) and (S.7), we see that $\widehat{\boldsymbol{B}}=\check{\boldsymbol{B}}$ and $\boldsymbol{D}_d\boldsymbol{V}_d\widehat{\boldsymbol{C}}_d=\check{\boldsymbol{C}}_d$, or equivalently, $\widehat{\boldsymbol{C}}_d=\boldsymbol{V}_d\boldsymbol{D}_d^{-1}\check{\boldsymbol{C}}_d$. □

S2 | ALGORITHM CONVERGENCE

Denote $g(\tilde{\boldsymbol{C}}_1,\ldots,\tilde{\boldsymbol{C}}_D,\boldsymbol{B})$ as the objective function from problem (15). Algorithm 1 implements a block coordinate descent scheme, where in the $(r+1)$'th iteration the conditional minimization problem

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

is solved for $d=1,\ldots,D$, and likewise for $\boldsymbol{B}^{(r+1)}$. In general, the convergence of such a scheme can be guaranteed if each of the sub-problems is convex and has a unique solution (Bertsekas, 1997). A minor augmentation can be made to the problem which guarantees convergence of the solution sequence, $\{\tilde{\boldsymbol{C}}_1^{(r)},\ldots,\tilde{\boldsymbol{C}}_D^{(r)},\boldsymbol{B}^{(r)}\}$, to a stationary point. In particular, adding a proximal regularization term to obtain the augmented problem

$$\min_{\boldsymbol{X}} g(\tilde{\boldsymbol{C}}_1^{(r+1)},\ldots,\tilde{\boldsymbol{C}}_{d-1}^{(r+1)},\boldsymbol{X},\tilde{\boldsymbol{C}}_{d+1}^{(r)},\ldots,\tilde{\boldsymbol{C}}_D^{(r)},\boldsymbol{B}^{(r)})+\frac{\mu_d^{(r)}}{2}\left\|\boldsymbol{X}-\tilde{\boldsymbol{C}}_d^{(r)}\right\|_F^2 \tag{S.8}$$

where $\mu_d^{(r)}>0$. This can be interpreted as an additional ridge penalty that "shrinks" the solution towards the value at the previous iteration. The augmented problem (S.8) is strongly convex and therefore has a unique solution. If we additionally assume boundedness of the solution sequence, (S.8) is guaranteed to converge to a stationary point of g. In practice, boundedness can be enforced by performing normalization to each of the matrices following each iteration of the algorithm. Note that due to the convergence of the solution sequence, the effect of the ridge penalty vanishes as $r \rightarrow \infty$.

Solving the augmented problem not only guarantees convergence but has also been shown to improve convergence speed (Li et al., 2013), particularly when $\mu_d^{(r)}$ is decreased over iterations. Razaviyayn et al. (2013) propose an empirically determined update rule

$$\mu_d^{(r+1)} = \mu_a + \mu_b \frac{\left\| \mathcal{G} - \sum_{k=1}^{K} \bigotimes_{d=1}^{D} \tilde{\boldsymbol{c}}_{d,k}^{(r)} \otimes \boldsymbol{b}_k^{(r)} \right\|_F}{\|\mathcal{G}\|_F}$$

for some constants small positive constants μ_a, μ_b.

S3 | MULTIDIMENSIONAL PENALIZED FPCA

In Section 2, we argued that the K-oMPB is an appealing option for modeling multidimensional functional data, both due to attractive theoretical properties and the existence of efficient numerical algorithms to form estimates in practice. That said, the efficiency of the approximation performance, provided in Theorem 3.3, depends on the joint low-rank structure of the eigenfunctions of U under the chosen marginal basis system. For many real world datasets, it may be the case that the first $K^* \ll K$ eigenfunctions can capture nearly the same proportion of variance of the data as the K-oMPB. Unfortunately, as was discussed in the introduction, the standard techniques for estimating the eigenfunctions (FPCA) suffer greatly from the curse of dimensionality. While K-oMPB may not be the optimal rank-K basis system, with an appeal to the low-rank property observed in many real world datasets it is reasonable to assume that for *most U* there exists some $K \ll m^D$, for which $\boldsymbol{\zeta}_m^*$ captures nearly the same proportion of variance as $\boldsymbol{\tau}_m$. Under this assumption, we can define a multidimensional FPCA using the *post-represented* data, i.e. the MPF representations of the sample $\mathcal{Y}$ obtained from Algorithm 1, which avoids the curse of dimensionality while incurring only trivial additional computational expense. In the remainder of this section we outline the procedure for the proposed two-stage FPCA.

Consider the method for FPCA proposed by Silverman (1996), in which the jth eigenfunction ψ_j is defined as the function maximizing the penalized sample variance with modified orthogonality constraints

$$\hat{\psi}_j = \max_{\psi \in \mathbb{W}^{\alpha,2}(\mathcal{M})} \frac{\sum_{i=1}^{N} \mathrm{Var}(\langle \psi, U_i \rangle_{\mathcal{H}})}{\langle \psi, \psi \rangle_\lambda} \tag{S.9}$$
$$\text{s.t.} \quad \|\psi\|_{\mathcal{H}}^2 = 1, \qquad \langle \psi, \psi_k \rangle_\lambda = 0, \text{ for } k = 1, 2, ..., j-1.$$

Here $\langle \psi, \psi_k \rangle_\lambda \equiv \langle \psi, \psi_k \rangle_{\mathcal{H}} + \lambda \langle L(\psi_j), L(\psi_k) \rangle_{\mathcal{H}}$ and $L : \mathbb{W}^{\alpha,2}(\mathcal{M}) \rightarrow \mathcal{H}$ is an αth order linear differential operator quantifying the global roughness. For simplicity, hereafter we define $L := \Delta_{\mathcal{M}}$, the Laplacian operator on $\mathcal{M}$, though other differential operators can be incorporated effortlessly. We choose a framework for FPCA that facilitates the optional incorporation of a flexible global roughness penalty in case a particular application requires a supplement to the marginally independent regularization imposed using penalties of the form in Definition **??**, e.g. penalizing mixed partial derivatives.

In the 1-dimensional case, the optimization problem (S.9) can be efficiently solved using a two-stage approach;

first computing $\widehat{U}_i$ 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: $\widehat{U}_i(\boldsymbol{x}) = \boldsymbol{b}_i'\boldsymbol{\zeta}^*(\boldsymbol{x})$, and then solve Equation (S.9) with the additional constraint $\hat{\psi}_j \in \text{span}(\boldsymbol{\zeta}^*)$, i.e. look for solutions $\hat{\psi}_j(\boldsymbol{x}) = \sum_{k=1}^{K} s_{jk}\zeta_k^*(\boldsymbol{x})$ for some $\boldsymbol{s}_j = (s_{j1}, ..., s_{jK})' \in \mathbb{R}^K$. Under this setup, the optimization problem (S.9) is equivalent to

$$\begin{aligned} \boldsymbol{s}_j &= \max_{\boldsymbol{s}} \frac{\boldsymbol{s}'\boldsymbol{J}_{\zeta^*}\boldsymbol{\Sigma}_{\boldsymbol{b}}\boldsymbol{J}_{\zeta^*}\boldsymbol{s}}{\boldsymbol{s}'\boldsymbol{J}_{\zeta^*}\boldsymbol{s} + \lambda \boldsymbol{s}'\boldsymbol{R}_{\zeta^*}\boldsymbol{s}} \\ \text{s.t.} \quad \boldsymbol{s}'\boldsymbol{J}_{\zeta^*}\boldsymbol{s} &= 1, \quad \boldsymbol{s}'[\boldsymbol{J}_{\zeta^*} + \lambda \boldsymbol{R}_{\zeta^*}]\boldsymbol{s}_k = 0, \text{ for } k = 1, 2, ..., j-1. \end{aligned} \tag{S.10}$$

Here $\boldsymbol{\Sigma}_{\boldsymbol{b}} = \text{Cov}(\boldsymbol{b})$ and $\boldsymbol{J}_{\zeta^*}$, $\boldsymbol{R}_{\zeta^*}$ are symmetric PSD matrices with elements $[\boldsymbol{J}_{\zeta^*}]_{ij} = \langle \zeta_i^*, \zeta_j^* \rangle_{\mathcal{H}}$ and $[\boldsymbol{R}_{\zeta^*}]_{ij} = \langle \Delta_{\mathcal{M}}(\zeta_i^*), \Delta_{\mathcal{M}}(\zeta_j^*) \rangle_{\mathcal{H}}$, respectively. The objective function in Equation (S.10) is a generalized Rayleigh quotient and it can be shown that the solutions for $j = 1, ..., K^*$ are equivalently defined by the first K^* solutions to the generalized eigenvalue problem

$$\boldsymbol{J}_{\zeta^*}\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{J}_{\zeta^*}\boldsymbol{s}_j = \gamma_j[\boldsymbol{J}_{\zeta^*} + \lambda \boldsymbol{R}_{\zeta^*}]\boldsymbol{s}_j. \tag{S.11}$$

Therefore, the vector of estimated eigenfunctions is $\hat{\boldsymbol{\psi}}(\boldsymbol{x}) := (\boldsymbol{s}_1'\boldsymbol{\zeta}^*(\boldsymbol{x}), ..., \boldsymbol{s}_{K^*}'\boldsymbol{\zeta}^*(\boldsymbol{x}))'$.

Notably, due to the marginal product structure of $\boldsymbol{\zeta}^*$, the D-dimensional integrals and partial derivatives required for the computation of $\boldsymbol{J}_{\zeta^*}$ and $\boldsymbol{R}_{\zeta^*}$ decompose into simple sums and products of integrals and derivatives over the marginal spaces. This highlights an important practical advantage of working with the marginal product structure, as it allows us to circumvent the potentially enormous computational cost of performing numerical integration/differentiation of an arbitrary D-dimensional function.

Of course, $\boldsymbol{\zeta}^*$, and hence $\boldsymbol{J}_{\zeta^*}$, $\boldsymbol{R}_{\zeta^*}$, and $\boldsymbol{\Sigma}_{\boldsymbol{b}}$ are unknown and must be estimated from the data. Algorithm S.1 provides pseudocode for performing the two stage regularized multidimensional FPCA. The algorithm requires the precomputation of the marginal inner product matrices defined by

$$\begin{aligned} \boldsymbol{J}_{\phi_d}(i,j) &= \langle \phi_{d,i}, \phi_{d,j} \rangle_{\mathcal{H}_d} \\ \boldsymbol{R}_{\phi_d}(i,j) &= \langle \Delta_{\mathcal{M}_d}(\phi_{d,i}), \Delta_{\mathcal{M}_d}(\phi_{d,j}) \rangle_{\mathcal{H}_d} \\ \boldsymbol{E}_{\phi_d}(i,j) &= \langle \phi_{d,i}, \Delta_{\mathcal{M}_d}(\phi_{d,j}) \rangle_{\mathcal{H}_d} \end{aligned}$$

Given the $\boldsymbol{C}_d$'s, simple derivations show that the marginal product structure of the ζ_k's permits fast analytic computation of $\boldsymbol{J}_{\zeta^*}$ and $\boldsymbol{R}_{\zeta^*}$ based on the element-wise formulas

$$\boldsymbol{J}_{\zeta^*}(i,j) = \prod_{d=1}^{D} \boldsymbol{c}_{d,i}'\boldsymbol{J}_{\phi_d}\boldsymbol{c}_{d,j} \tag{S.12}$$

and

$$\begin{aligned} \boldsymbol{R}_{\zeta^*}(i,j) = &\sum_{d=1}^{D}(\prod_{b \neq d}^{D} \boldsymbol{c}_{b,i}'\boldsymbol{J}_{\phi_b}\boldsymbol{c}_{b,j})\boldsymbol{c}_{d,i}'\boldsymbol{R}_{\phi_d}\boldsymbol{c}_{d,j} \\ &+ \sum_{\substack{a,d \\ a \neq d}}(\prod_{\substack{b \neq a \\ b \neq d}}^{D} \boldsymbol{c}_{b,i}'\boldsymbol{J}_{\phi_b}\boldsymbol{c}_{b,j})(\boldsymbol{c}_{d,i}'\boldsymbol{E}_{\phi_d}\boldsymbol{c}_{d,j})(\boldsymbol{c}_{a,i}'\boldsymbol{E}_{\phi_a}\boldsymbol{c}_{a,j}). \end{aligned} \tag{S.13}$$

In practice, we form estimates $\widehat{\boldsymbol{C}}_d$ using Algorithm 1. The corresponding estimates of the inner product matrices $\widehat{\boldsymbol{J}}_{\zeta^*}$ and $\widehat{\boldsymbol{R}}_{\zeta^*}$ are obtained by plugging $\widehat{\boldsymbol{C}}_d$ into (S.12) and (S.13), respectively.

We offer a couple quick remarks on the practical implementation of Algorithm S.1. There are several ways to solve the generalized eigenvalue problem (S.11). We used Algorithm 9.4.2 in Ramsay and Silverman (2005), since it avoids direct construction of $\boldsymbol{\Sigma}_b$ which is important for numerical stability in the high dimensional case when $K > N$. If subsequent dimensionality reduction is desirable, an initial K-MPB with a very large K can be estimated and then used to construct the basis system $\{\psi_1, ..., \psi_{K^*}\}$, where $K^* < K$ can be chosen by a threshold on the desired proportion of variance explained. In practice, it is often the case that a $K^* \ll K$ can explain a large proportion of the functional variance.

Algorithm 1 Two-Stage Regularized Multidimensional FPCA

```
Estimate $\{\tilde{\boldsymbol{C}}_d\}$, $\widehat{\boldsymbol{B}}$ using Algorithm 1
Transform $\widehat{\boldsymbol{C}}_d = \boldsymbol{V}_d'\boldsymbol{D}_d^{-1}\tilde{\boldsymbol{C}}_d$ for $d = 1, ..., D$
Assemble $\widehat{\boldsymbol{J}}_{\zeta^*}$ and $\widehat{\boldsymbol{R}}_{\zeta^*}$ using (S.12) and (S.13), respectively.
Compute $\widehat{\boldsymbol{S}} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_{K^*}]$ using Algorithm 9.4.2 in Ramsay and Silverman (2005).
```

S4 | DATA ADAPTIVE MARGINAL BASIS

In the following, we develop an alternative framework to estimating $\xi_{k,d}$ over a prespecified ϕ_d in which the marginal basis systems are allowed to be data adaptive.

S4.1 | Overview of Random Projections

Our construction of a data adaptive marginal basis system is built off of the framework of random projections. The utility of random projection techniques for performing decomposition of large matrices is elaborated in the influential work of Halko et al. (2011). Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, where m and n are large, random projection methods aim to construct an optimal rank $r \ll \min(m, n)$ subspace, spanned by the columnwise orthogonal $\boldsymbol{U} \in \mathbb{R}^{m \times r}$, such that

$$\boldsymbol{A} \approx \boldsymbol{U}\boldsymbol{U}'\boldsymbol{A}.$$

Whatever downstream matrix decomposition is of interest can instead be approximated by applying it to the projected matrix $\boldsymbol{U}'\boldsymbol{A}$, at a much lower computational cost.

So how is $\boldsymbol{U}$ formed? This is where randomness becomes useful. Define the random vector $\boldsymbol{z} \in \mathbb{R}^n$ whose elements are, for example, $i.i.d.$ standard normal. Forming a sample of r realizations of $\boldsymbol{z}$, randomness ensures with high probability that the set of sample vectors $\{\boldsymbol{z}_i : i = 1, ..., r\}$ are both in general linear position and linearly independent. As a result, the set of transformed random vectors $\{\boldsymbol{A}\boldsymbol{z}_i : i = 1, ..., r\}$ are both non-zero and linearly independent. If the rank of $\boldsymbol{A}$ is exactly r, then the set of transformed vectors will span its columnspace with high probability and therefore $\boldsymbol{U}$ can be constructed by forming an orthogonal basis for the set, e.g. by the QR-decomposition. Rather than having exactly rank r, it is often more reasonable to assume that $\boldsymbol{A}$ can be decomposed into the sum of a rank r matrix (the signal) and a residual matrix with relatively small expected norm (the noise). In this case, it can actually be beneficial to draw more than r samples of $\boldsymbol{z}$ in order to fully interrogate the columnspace of the low rank signal, due to the effect of the residual component (Halko et al., 2011).

S4.2 | Reformulating the Loss

Consider the objective function of the optimization problem (15). Notice that the loss term depends on ϕ_d only through $\boldsymbol{U}_d$, i.e. in the construction of $\widehat{\mathcal{G}}$. Therefore, we can recast the problem of selecting a set of optimal marginal basis systems as the problem of selecting a set of $\boldsymbol{U}_d$ that, in some sense, produce an optimal compression of $\mathcal{Y}$. We propose to form $\boldsymbol{U}_d$ using random projections applied to the corresponding d-mode matricization $\mathcal{Y}$.

The proposed procedure is outlined in Algorithm S.2, which is essentially a combination of Algorithms 4.2 and 4.4 from Halko et al. (2011). The first for loop implements power iterations in order to increase the decay rate of the spectrum of the $\boldsymbol{Y}_{(d)}$ while retaining the same singular vectors. Data adaptive selection of the marginal ranks m_d is then performed in the while loop and is guided by the tuning parameter tol, which ensures m_d is large enough that

$$\|(\boldsymbol{I} - \boldsymbol{U}_d\boldsymbol{U}_d')\boldsymbol{Y}_{(d)}\|_F < \text{tol}.$$

with probability at least $1 - 10^{-r}$ (Halko et al., 2011).

Algorithm 2 Data-driven $\boldsymbol{U}_d$ via Random Projections

```
Input $\boldsymbol{Y}_{(d)}$, tol, q, r
Output $\boldsymbol{U}_d$
Initialize $\boldsymbol{U}_d$ as an empty $n_d \times 0$ matrix, $j = 1$
Sample $n_d$ $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ of length $N\prod_{j\neq d} n_j$
$\boldsymbol{A}_d = \boldsymbol{Y}_{(d)}[\boldsymbol{z}_1, \dots, \boldsymbol{z}_d]$
for l=1,...,q do
    From the LU-decomposition $\boldsymbol{L}_{lower}, \boldsymbol{U}_{upper} = \text{lu}(\boldsymbol{A}_d)$
    Form the LU-decomposition of the product $\boldsymbol{L}_{lower}, \boldsymbol{U}_{upper} = \text{lu}(\boldsymbol{Y}_{(d)}'\boldsymbol{L}_{lower})$
    $\boldsymbol{A}_d = \boldsymbol{Y}_{(d)}\boldsymbol{L}_{lower}$
while max$\{\|A_d(:,j)\|, \dots, \|A_d(:,j+r-1)\|\} \geq$ tol$\|\boldsymbol{A}_d\|_F/(10\sqrt{2\pi})$ do
    $\boldsymbol{A}_d(:,j) = [\boldsymbol{I} - \boldsymbol{U}_d(:,1:j-1)\boldsymbol{U}_d'(:,1:j-1)]\boldsymbol{A}_d(:,j)$
    $\boldsymbol{U}_d(:,j) = \boldsymbol{A}_d(:,j)/\|\boldsymbol{A}_d(:,j)\|$
    $\boldsymbol{A}_d(:,j+r) = [\boldsymbol{I} - \boldsymbol{U}_d(:,1:j)\boldsymbol{U}_d'(:,1:j)]\boldsymbol{A}_d(:,j+r)$
    for l=j+1,...,j+r-1 do
        $\boldsymbol{A}_d(:,l) = \boldsymbol{A}_d(:,l) - \langle \boldsymbol{A}_d(:,l), \boldsymbol{U}_d(:,j)\rangle \boldsymbol{U}_d(:,j)$
    $j = j + 1$
```

We note that a variety of randomized methods have been developed to perfrom tensor decomposition, including subsampling (Battaglino et al., 2018) and randomized projection techniques (Erichson et al., 2020). Algorithm S.2 is similar to Algorithm 1 from Erichson et al. (2020), except the latter prespecifies a single parameter for all of the marginal ranks and uses oversampling to accommodate for noise, as opposed to the proposed scheme which permits the data adaptive selection of marginal ranks controlled by a relative error tolerance parameter.

S4.3 | Regularization

Although the loss term in (15) can be formulated in a manner agnostic to ϕ_d, the roughness penalty depends on the smoothness of the candidate estimates, quantified by $\boldsymbol{T}_d$ which, under the construction given in the proof to Proposition 2.2, depends explicitly on ϕ_d. We now describe a technique to estimate $\boldsymbol{T}_d$ without explicitly specifying ϕ_d.

Let Q_d be a quadrature rule with

$$\int_{\mathcal{M}_d} h(x_d) = \sum_{l=1}^{q_d} w_{d,l} h(x_{d,l}) + E(h),$$

where $\{(w_{d,l}, x_{d,l})\}$ are the weights and points defining the quadrature rule, h is the function to be integrated and $E(h)$ is the error term of the approximation, which depends on both Q_d and h. Let $\boldsymbol{L}_{d,Q} \in \mathbb{R}^{q_d \times n_d}$ be the estimates of differential operator $L_d(h)$ at the quadrature points in Q_d based on the point evaluations of h over $\boldsymbol{x}_d$. In the simplest case, the points in Q_d are the same as the equispaced marginal grid points and then $\boldsymbol{L}_{d,Q}$ may be formed using finite differences with the appropriate stencil over $\boldsymbol{x}_d$. In general, it may be desirable to specify a quadrature rule over a different grid, for example if $\boldsymbol{x}_d$ is very dense, in which case (linear) imputation would be necessary to form the desired estimates at the quadrature points. In any case, we can form an estimate for the marginal roughness penalty by plugging in the following estimator of $\boldsymbol{T}_d$

$$\widehat{\boldsymbol{T}}_d = \boldsymbol{U}_d' \boldsymbol{L}_{d,Q}' \boldsymbol{Q}_d \boldsymbol{L}_{d,Q} \boldsymbol{U}_d, \tag{S.14}$$

where $\boldsymbol{Q}_d = \text{Diag}(w_{d,1}, ..., w_{d,q_d})$ is the diagonal matrix of numerical quadrature points.

S4.4 | Obtaining a Continuous Representation

In the proposed adaptive marginal basis methodology, instead of specifying ϕ_d directly, we form a data driven estimate of $\boldsymbol{U}_d$ using Algorithm S.2 along with the discretized estimate of $\boldsymbol{T}_d$ using the method outlined in Section S4.3. These quantities can then be used in Algorithm 1 to construct $\tilde{\boldsymbol{C}}_d$. In the context of our model, recall that

$$\Xi_d = \boldsymbol{\Phi}_d \boldsymbol{C}_d = \boldsymbol{U}_d \tilde{\boldsymbol{C}}_d,$$

therefore we can form estimates of the optimal marginal basis function evaluations at each of the marginal grid points. To obtain the final continuous representations of $\xi_{d,k}$, a marginal function approximation, e.g. smoothing or interpolation, is performed.

S5 | ADDITIONAL SIMULATION STUDIES

S5.1 | Overview of Sandwich Smoother

The so-called *sandwich smoother*, first introduced by Xiao et al. (2013) and then extended by French and Kokoszka (2021), is a method for estimating the coefficients of a tensor product approximation to an unknown deterministic function from noisy observations on a grid. The main contribution is in a clever formulation of the penalty term, which allows for the fast computation of the GCV statistic and hence a computationally efficient technique for selecting the roughness penalty strength. For more information, see either of the aforementioned papers or the `hero` package in R (French, 2020).

S5.2 | Overview of FCP-TPA

We give a brief overview of the the FCP-TPA algorithm (Allen, 2013), which is essentially a D-dimensional extension of the 2-dimensional regularization scheme from Huang et al. (2009). As such, it does not consider the same problem that the current work explores, namely optimal basis systems for representing a random sample of functional data. Rather they consider the canonical nonparametric function approximation task, in which there is a single, unknown deterministic function discretely observed with error. That said, it is easy enough to adapt their method to our situation, i.e. by introducing a non-smooth “subject mode”. This is what is presented below.

Using our notation, the FCP-TPA estimates the kth *MPF basis evaluation vectors* $\Xi_{d,k}$ and associated coefficient vector $\boldsymbol{b}_k$ by solving a series of K rank-one penalized decompositions of the residual tensor. That is, at the kth iteration, FCP-TPA solves problem

$$\min_{\Xi_{1,k},\ldots,\Xi_{D,K},\boldsymbol{b}_k} \left\| \mathcal{Y}_{resid} - \bigotimes_{d=1}^{D} \Xi_{d,k} \otimes \boldsymbol{b}_k \right\|_F^2 - \prod_{d=1}^{D} \left\| \Xi_{d,k} \right\|_2^2 + \prod_{d=1}^{D} \Xi'_{d,k} \boldsymbol{P}_d^{-1} \Xi_{d,k} \tag{S.15}$$

where $\mathcal{Y}_{resid} = \mathcal{Y} - \sum_{j=1}^{k-1} \bigotimes \Xi_{d,j} \otimes \boldsymbol{b}_j$ and $\boldsymbol{P}_d \in \mathbb{R}^{n_d \times n_d}$ is a smoothing matrix, e.g. derived using squared second order differences. The solution to (S.15) is approximated using tensor power iterations, which are shown to converge to a stationary point.

Note that FCP-TPA does not directly construct a continuous representation but rather the discrete evaluations of the optimal marginal product functions on the observed marginal grid, i.e. the Ξ_d's. In order to obtain a continuous representation from the output of FCP-TPA, the decompose-then-represent approach is used in which the marginal basis functions are estimated from the basis expansion of the Ξ_d's.

S5.3 | Representing Random Marginal Product Functions

Table S1 displays a comparison of the performance of the K-oMPB estimated by Algorithm 1 (denoted `MARGARITA`) and the tensor product basis estimated by the sandwhich smoother. For all combinations of distribution and sampling plans considered, `MARGARITA` exhibits lower moMISE over the sample than the tensor product basis when comparing models with similar number degrees of freedom. `MARGARITA` is often able to produce better fits than the tensor product basis with substantially fewer degrees of freedom. For example, the K-oMPB with 675 degrees of freedom exhibits lower moMISE than the tensor product basis system with 2, 197 degrees of freedom for all considered settings.

The strong performance of the `MARGARITA` is driven at least in part by the ability to share information across the sample to estimate a set of basis functions that conforms to the distribution. From table S1, we can see this is of particular importance in the case of low SNR, for which access to a larger sample size, $N = 50$ vs. $N = 5$, typically results in significantly lower moMISE. A caveat here is that for the gain from information pooling to be realized, the rank of the model must be “large enough”, e.g. compare results from $K_{fit} = 8$ to the $K_{fit} = 15, 25$ case. This is a distinct advantage over the tensor product basis system, whose estimation is done for each subject independently.

Table S2 displays the relative difference in moMISE, defined as

$$\frac{\text{moMISE}_{\text{FCP-TPA}} - \text{moMISE}_{\texttt{MARGARITA}}}{\text{moMISE}_{\text{FCP-TPA}}} \tag{S.16}$$

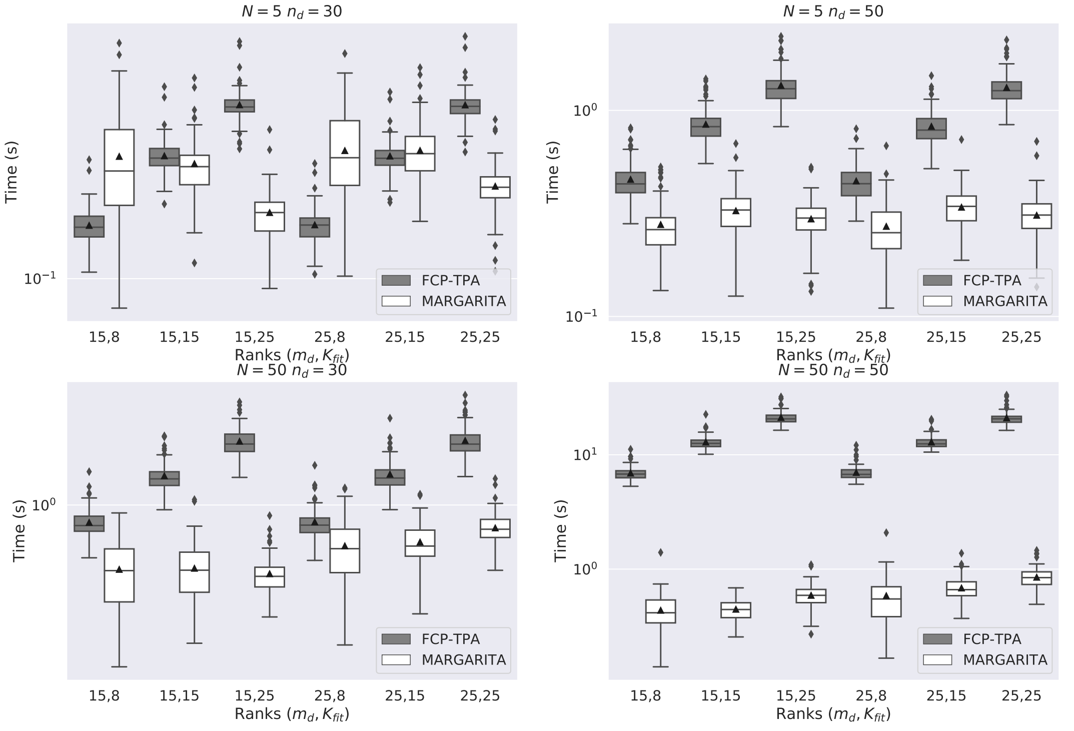

FIGURE S1 Computational time comparison between FCP-TPA and MARGARITA. The Y-axis is plotted on log-scale for clarity.

between the fits resulting from the FCP-TPA algorithm and MARGARITA. As all but one of the entries in the table is positive, MARGARITA is nearly uniformly outperforming the FCP-TPA. The one configuration for which FCP-TPA has lower moMISE is the low SNR, small sample size along with large marginal rank $m_d = 25$, which is somewhat pathological and both methods perform poorly. Echoing the results presented in the main text, we see the boost in performance from MARGARITA generally increases with K_{fit}.

Figure S1 shows a computational time comparison between the two algorithms. For all panels, we fixed $K_t = 20$ and $\sigma^2 = 10.0$, but similar patterns emerge for the other cases. For most cases MARGARITA is faster than FCP-TPA, the only exceptions being for very small sample size and small ranks, which may be considered as toy examples.

TABLE S1 moMISE comparison of `MARGARITA` (a), to the tensor product basis estimated by sandwhich smoother (b).

(a) Marginal Product Basis

				K_{fit} = 15			K_{fit} = 25		
				m_d (model d.o.f.)					
K_{true}	σ^2	N	n_d	8 (360)	15 (675)	25 (1,125)	8 (600)	15 (1,125)	25 (1,875)
10	0.5	5	30	0.0890	0.0433	0.0433	0.0511	0.0030	0.0033
10	0.5	5	50	0.0932	0.0405	0.0402	0.0560	0.0006	0.0006
10	0.5	50	30	0.1186	0.0426	0.0411	0.0790	0.0017	0.0007
10	0.5	50	50	0.1196	0.0398	0.0394	0.0757	0.0001	0.0001
10	10.0	5	30	0.1148	0.1049	0.1634	0.1039	0.0919	0.1579
10	10.0	5	50	0.0976	0.0530	0.0647	0.0616	0.0180	0.0312
10	10.0	50	30	0.1166	0.0525	0.0632	0.0800	0.0146	0.0263
10	10.0	50	50	0.1149	0.0423	0.0441	0.0745	0.0019	0.0033
20	0.5	5	30	0.4910	0.1158	0.0684	0.4418	0.0511	0.0059
20	0.5	5	50	0.4872	0.1075	0.0627	0.4429	0.0475	0.0015
20	0.5	50	30	0.6334	0.1239	0.0679	0.5850	0.0564	0.0025
20	0.5	50	50	0.6388	0.1182	0.0646	0.5998	0.0539	0.0009
20	10.0	5	30	0.5058	0.1587	0.1723	0.4736	0.1200	0.1492
20	10.0	5	50	0.4994	0.1185	0.0841	0.4534	0.0594	0.0286
20	10.0	50	30	0.6544	0.1353	0.0867	0.6008	0.0682	0.0220
20	10.0	50	50	0.6415	0.1231	0.0674	0.5915	0.0579	0.0039

(b) Tensor Product Basis

				m_d (model d.o.f.)					
K_{true}	σ^2	N	n_d	7 (343)	8 (512)	9 (729)	11 (1,331)	12 (1,728)	13 (2,197)
10	0.5	5	30	1.1409	0.9520	0.9096	0.5327	0.4817	0.2136
10	0.5	5	50	1.1251	0.9326	0.8870	0.5067	0.4581	0.1823
10	0.5	50	30	1.1457	0.9556	0.9130	0.5338	0.4840	0.2134
10	0.5	50	50	1.1170	0.9251	0.8789	0.5020	0.4516	0.1832
10	10.0	5	30	1.1579	1.0224	1.0136	0.7705	0.7559	0.6051
10	10.0	5	50	1.1476	0.9677	0.9300	0.5822	0.5492	0.3027
10	10.0	50	30	1.1822	1.0465	1.0369	0.7861	0.7721	0.6109
10	10.0	50	50	1.1518	0.9679	0.9334	0.5857	0.5517	0.3036
20	0.5	5	30	1.8024	1.6254	1.4738	0.9819	0.7065	0.3806
20	0.5	5	50	1.7445	1.5693	1.4194	0.9342	0.6623	0.3464
20	0.5	50	30	1.8352	1.6530	1.4973	1.0030	0.7218	0.3842
20	0.5	50	50	1.8207	1.6349	1.4739	0.9748	0.6911	0.3580
20	10.0	5	30	1.8076	1.6757	1.5688	1.2221	1.0375	0.8092
20	10.0	5	50	1.8592	1.6807	1.5290	1.0597	0.7971	0.4817
20	10.0	50	30	1.9789	1.8289	1.7073	1.3210	1.1113	0.8472
20	10.0	50	50	1.8527	1.6765	1.5280	1.0560	0.7946	0.4833

TABLE S2 Relative difference in moMISE, as defined by equation (S.16), for FCP-TPA (Allen, 2013) and `MARGARITA` for marginal ranks 15 and 25 and $K_{fit} = 8, 15, 25$. Positive values indicate lower moMISE for `MARGARITA`. A grid search to select λ_d was performed for each fit and the results from the optimal value are reported. The entry in bold face indicates the **only case** that FCP-TPA outperformed `MARGARITA`.

				$K_{fit} = 15$			$K_{fit} = 25$		
				m_d					
K_{true}	σ^2	N	n_d	8	15	25	8	15	25
10	0.5	5	30	0.2113	0.2506	0.1738	0.3062	0.7949	0.7160
10	0.5	5	50	0.2544	0.2433	0.1534	0.3634	0.9328	0.8518
10	0.5	50	30	0.2062	0.2118	0.1555	0.2323	0.9228	0.8734
10	0.5	50	50	0.1597	0.1850	0.1167	0.2954	0.9757	0.9437
10	10.0	5	30	0.3684	0.2533	**-0.0770**	0.4413	0.4426	0.4895
10	10.0	5	50	0.3086	0.3824	0.2120	0.4409	0.6580	0.4486
10	10.0	50	30	0.3487	0.3509	0.1540	0.3529	0.6836	0.5607
10	10.0	50	50	0.2080	0.3300	0.2446	0.3079	0.7229	0.7629
20	0.5	5	30	0.0946	0.4304	0.4429	0.1152	0.6157	0.8706
20	0.5	5	50	0.1139	0.4149	0.4505	0.1243	0.6088	0.9428
20	0.5	50	30	0.0708	0.4452	0.4624	0.0743	0.6107	0.9699
20	0.5	50	50	0.0835	0.4274	0.4554	0.0880	0.6092	0.9816
20	10.0	5	30	0.1235	0.4651	0.3642	0.1385	0.5306	0.6020
20	10.0	5	50	0.1329	0.4600	0.5117	0.1382	0.5928	0.7476
20	10.0	50	30	0.0954	0.4489	0.5313	0.0753	0.5856	0.8161
20	10.0	50	50	0.0915	0.4349	0.5089	0.0983	0.6155	0.9299

TABLE S3 Monte Carlo average MISE for representing a new realization for both `MARGARITA` and `MargFPCA`, for a variety of ranks and training sample sizes.

	MARGARITA				**MargFPCA**			
N_{train}	10	20	50	100	10	20	50	100
5	1.5880	1.4188	1.3746	1.3252	2.3361	2.1493	2.2462	2.1207
10	0.8895	0.7456	0.6930	0.6961	1.9747	1.9264	1.8646	1.9286
15	0.5036	0.3945	0.3203	0.3024	1.7118	1.7337	1.5499	1.5952
20	0.2898	0.1980	0.1505	0.1280	1.4634	1.3581	1.3520	1.2920
30	0.1393	0.0517	0.0313	0.0274	1.1477	1.0366	0.9847	1.0073

S5.4 | Generalization Performance

Figure S2 and Table S3 display the results of the simulation study in Section 4.2 for a variety of training sample sizes and ranks. Figure S3 displays the estimates of the first three eigenfunctions using the two stage FPCA approach from Section S3, using `MARGARITA` with $K = 60$ and $N_{train} = 200$, along with the ground truth.

S5.5 | Deterministic Functions

In this section, we briefly demonstrate the proposed methods utility for a different but related task, that of approximating the output of a deterministic analytic test function. The proposed framework is easily adapted to this situation by setting the subject mode dimension to 1. We evaluate our method on the benchmark Friedman 2 and 3 functions, which have the analytic form

$$y_i = \sqrt{x_{1i}^2 + \left(x_{2i}x_{3i} - \frac{1}{x_{2i}x_{4i}}\right)^2}$$

and

$$y_i = \arctan\left(\frac{x_{2i}x_{3i} - (x_{2i}x_{4i})^{-1}}{x_{1i}}\right)$$

respectively, where $\mathcal{M} = [0, 100]\times[40\pi, 560\pi]\times[0, 1]\times[0, 11]$. A total of $n_1 = n_2 = n_2 = n_4 = 20$ equispaced observations are sampled in each marginal dimension. A marginal basis system of cubic B-splines is selected for modeling.

For the Friedman 2 function, with only 5 marginal basis functions in each dimension, a rank-1 marginal product basis model is able to obtain an $r^2 \approx 0.998$. Using the same set-up for Friedman 3 function approximation, the resulting model fit has an $r^2 \approx 0.813$. We are able to increase this $r^2 > 0.99$ by doubling the number of marginal cubic B-splines to 10 and increasing the rank to 5.

We also consider the performance of the marginal product basis estimated by `MARGARITA` on a standard nonparametric regression task, with deterministic but unknown regression function and additive Gaussian errors. In the context of the current work, this task may correspond to estimating the mean function of U from the pooled data tensor. Specifically,

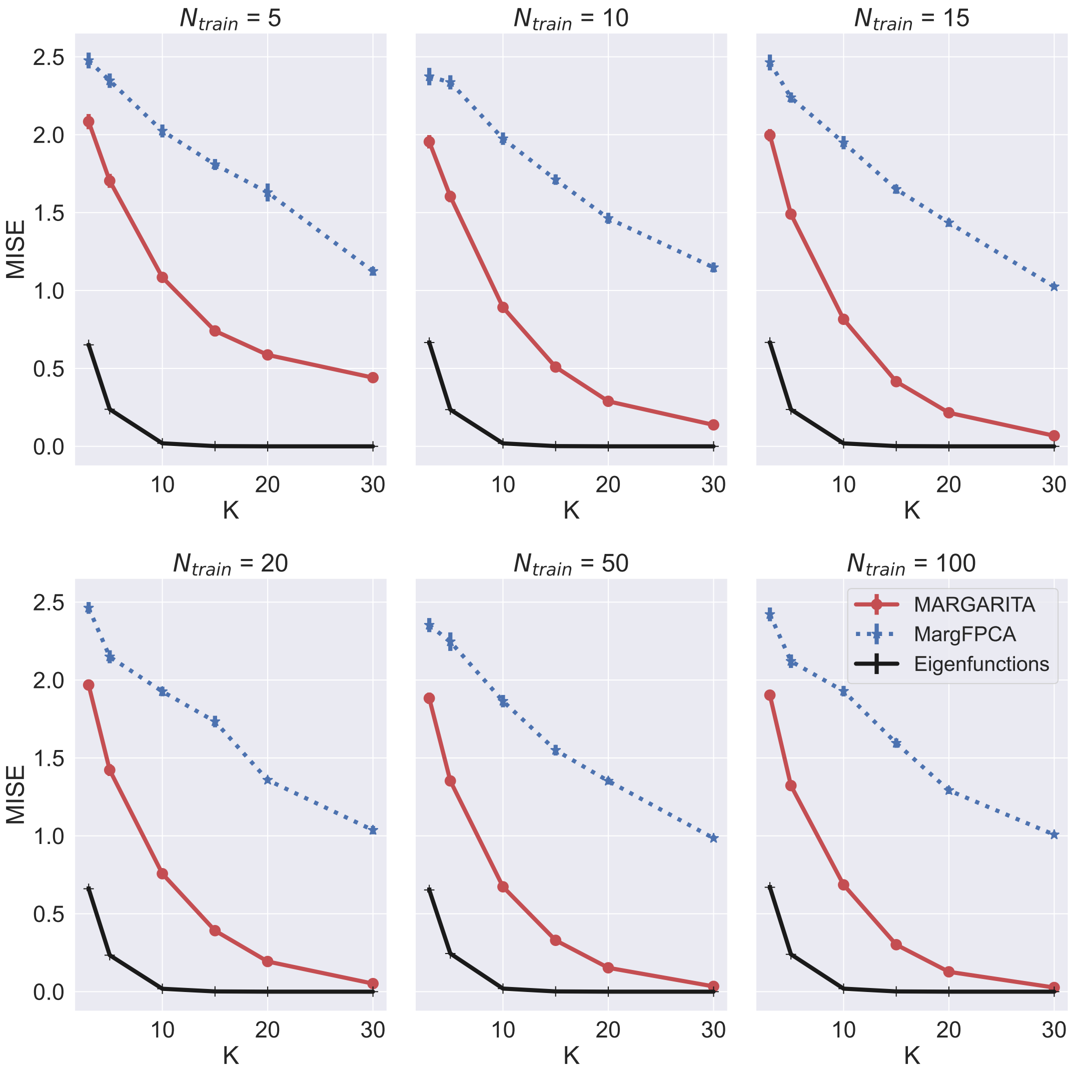

FIGURE S2 Comparison of the generalization performance of `MARGARITA` (red-solid) and `MargFPCA` (blue-dotted), for a variety of ranks and training sample sizes. The true optimal rank K basis system, i.e. the eigenfunctions, are included for comparison (black solid).

we use the observation model

$$y_i = \gamma\left(x_{2i} - x_{1i}\left(\frac{1}{3}x_{1i}^2 - 3\right)\right) + \epsilon_i,$$

where γ is the standard normal density function. The performance of `MARGARITA` is compared with tensor product basis estimation, from here on abbreviated TPB, and generalized additive model, denoted GAM. In order to facilitate fair comparison between the models, we enforce the total number of parameters to be roughly equivalent. To this

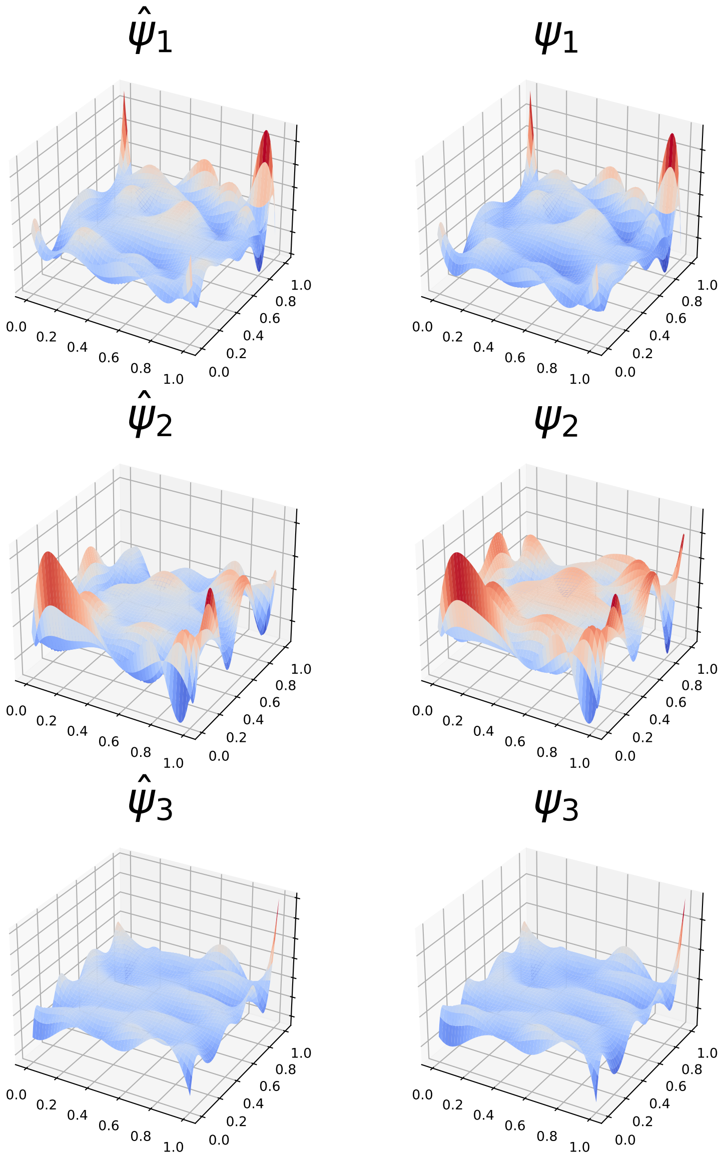

FIGURE S3 Two-stage eigenfunction estimates (left column) vs. ground truth (right column).

end, for `MARGARITA` K is set to 5 and cubic B-splines of rank 18 are used for the marginal basis systems, giving a total of 180 parameters to estimate. Cubic B-splines are also used for the TPB and GAM models, with 14 in each marginal direction for the TPB (196 total parameters) and 90 in each dimension for the GAM (180 total parameters). TPB is estimated via the sandwhich smoothing technique from Xiao et al. (2013), implemented in the hero package (French, 2020). The GAM is estimated using the `mgcv` package (Wood, 2011). The smoothing parameters for both the TPB and GAM models were chosen using a GCV criteria, while a grid search is performed to select the marginal smoothing

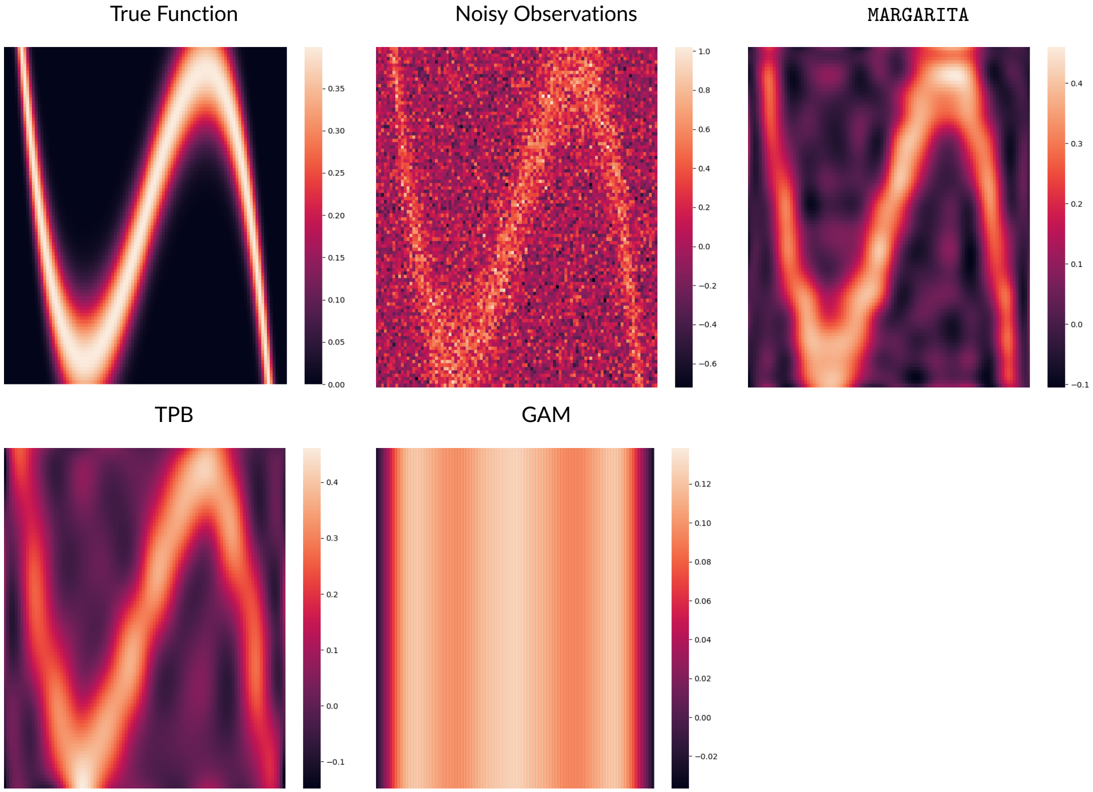

FIGURE S4 Comparison of `MARGARITA`, TPB and GAM fits for the canonical non-parametric regression task, for a highly non-linear unknown function.

parameters in `MARGARITA`. The domain is set to be $\mathcal{M} = [-4, 4]^2$ and samples were taken on a regular grid constructed from 100 equispaced points in each marginal direction.

Figure S4 shows heatmaps of the true function, noisy observations and fits from each of the models considered. The `MARGARITA` fit resulted in the lowest root mean squared error (RMSE) from the true data, ≈ 0.045 , followed by the TPB fit with RMSE ≈ 0.053 and finally the GAM fit with ≈ 0.14. The GAM struggles due to the non-additive (in x_1, x_2) nature of the regression function, only being able to model large scale features such as the true function being near to 0 at $x_1 = \pm 4$. Adding the "interaction term" $x_1 x_2$ to the GAM model did not improve the fit substantially. Upon visual inspection, `MARGARITA` and TPB produce similar quality fits in the rectangular region of $-2 \leq x_1 \leq 2$. `MARGARITA` does appear to better reconstruct the features in the boundary region near the corners $(4, -4), (-4, 4)$.

S6 | REAL DATA ANALYSIS

All subjects in our study were referred to the University of Rochester Medical Imaging Center and imaged on the same 3T MRI scanner. Study inclusion criteria included history of concussion, while exclusion criteria included dental braces, prior brain surgery, ventricular shunt, skull fractures, or other standard contraindications for MR imaging.

Diagnosis of concussion was made by neurologists, physical medicine and rehabilitation physicians, and sports medicine physicians. The control group consisted of young athletes with no history of concussion. The University Institutional Review Board approved this retrospective study. All MRI examinations were reviewed by an experienced neuroradiologist for any artifacts that might affect the quality of the study, as well as for the presence of recent or remote intracranial hemorrhage, signal abnormalities in the brain, hydrocephalus, congenital or developmental anomalies.

The diffusion MRI data was collected on a single 3T scanner using a 20-channel head coil (Siemens Skyra, Erlangen, Germany). Diffusion imaging was performed with a b-value of $1000\frac{s}{mm^2}$, using 64 diffusion-encoding directions. In addition, a $b = 0\frac{s}{mm^2}$ image was collected for signal normalization. Additional dMRI parameters included: $FOV = 256{\times}256mm$, number of slices = 70, image resolution = $2{\times}2{\times}2mm^3$, $TR/TE = 9000/88ms$, Generalized autocalibrating partially parallel acquisition (GRAPPA) factor $= 2$. Acquisition of dMRI data took 10 minutes and 14 seconds. A Gradient-recalled echo (GRE) sequence was also collected with $TEs = 4.92, 7.38ms$ at the same resolution of the dMRI to correct for susceptibility-induced distortion effects.

A diffusion tensor model (DTI) was fit to each subject's diffusion data and used to compute the per-voxel FA. Registration of the FA images to to the ICBM 2009c Nonlinear Symmetric 1mm template (Fonov et al., 2009) was then performed using the popular ANTS software (Avants et al., 2009). The domain of analysis was constrained to be the convex hull of a rectangular $115 \times 140 \times 120$ voxel grid in the template space covering the white matter, i.e. the raw data tensor $\mathcal{Y} \in \mathbb{R}^{115\times140\times120\times50}$. A white matter mask was also applied to the aligned data.

Figure S5 shows a pair plot of the coefficients associated with 3 most informative eigenfunctions, as measured by the magnitude of the associated regression coefficient in the logistic classifier, for one of the predictive models trained during cross validation. Notice that there exists a substantial degree of separation for several of the bivariate plots.

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.

Avants, B. B., Tustison, N. and Song, G. (2009) Advanced normalization tools (ants). *Insight j*, **2**, 1–35.

Battaglino, C., Ballard, G. and Kolda, T. G. (2018) A practical randomized CP tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, **39**, 876–901. URL: `https://doi.org/10.1137/17M1112303`.

Bertsekas, D. P. (1997) Nonlinear programming. *Journal of the Operational Research Society*, **48**, 334–334. URL: `https://doi.org/10.1057/palgrave.jors.2600425`.

Erichson, N. B., Manohar, K., Brunton, S. L. and Kutz, J. N. (2020) Randomized CP tensor decomposition. *Machine Learning: Science and Technology*, **1**, 025012. URL: `https://doi.org/10.1088/2632-2153/ab8240`.

Fonov, V., Evans, A., McKinstry, R., Almli, C. and Collins, D. (2009) Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, **47**, S102.

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

French, J. P. and Kokoszka, P. S. (2021) A sandwich smoother for spatio-temporal functional data. *Spatial Statistics*, **42**, 100413. URL: `https://www.sciencedirect.com/science/article/pii/S2211675320300075`. Towards Spatial Data Science.

Halko, N., Martinsson, P. G. and Tropp, J. A. (2011) Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, **53**, 217–288. URL: `https://doi.org/10.1137/090771806`.

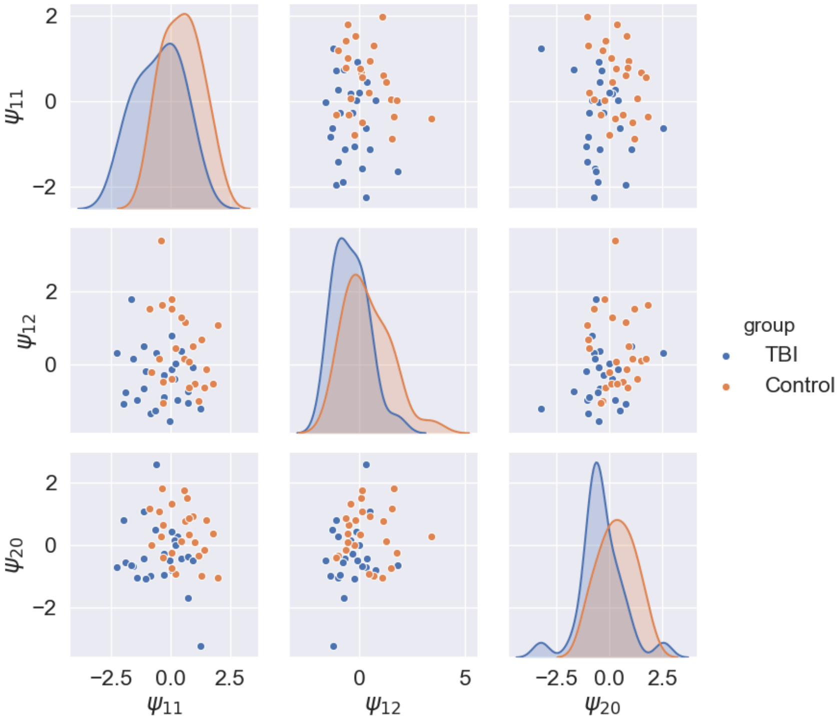

FIGURE S5 Pairwise scatter plots of the coefficients associated with the top 3 most important eigenfunctions, as determined by the logistic model. The panels along the diagonal show KDE for the marginal distributions. Each multidimensional eigenfunction provides some discriminatory power, as shown in the separation in the biplots. The logistic model trained on the full set of eigenfunction coefficients is able to distinguish TBI subjects from control with nearly 90% cross validated accuracy.

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.

Li, N., Kindermann, S. and Navasca, C. (2013) Some convergence results on the regularized alternating least-squares method for tensor decomposition. *Linear Algebra and its Applications*, **438**, 796–812. URL: https://www.sciencedirect.com/science/article/pii/S0024379511007828.

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

Razaviyayn, M., Hong, M. and Luo, Z.-Q. (2013) A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, **23**, 1126–1153. URL: https://doi.org/10.1137/120891009.

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.

van der Vaart, A. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*.

Vaart, A. W. v. d. (1998) *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wood, S. N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, **73**, 3–36.

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`.