Received: 24 January 2020 | Revised: 14 July 2020 | Accepted: 15 September 2020

DOI: 10.1111/biom.13385

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC METHODOLOGY

Modeling sparse longitudinal data on Riemannian manifolds

Xiongtao Dai[1] | **Zhenhua Lin**[2] | **Hans-Georg Müller**[3]

[1] Department of Statistics, Iowa State University, Ames, Iowa

[2] Department of Statistics and Applied Probability, National University of Singapore, Singapore

[3] Department of Statistics, University of California, Davis, California

Correspondence
Xiongtao Dai, Department of Statistics, Iowa State University, 2438 Osborn Dr, Ames, IA 50011.
Email: xdai@iastate.edu

Funding information
Division of Mathematical Sciences, Grant/Award Numbers: DMS-1712864, DMS-2014626

Abstract

Modern data collection often entails longitudinal repeated measurements that assume values on a Riemannian manifold. Analyzing such longitudinal Riemannian data is challenging, because of both the sparsity of the observations and the nonlinear manifold constraint. Addressing this challenge, we propose an intrinsic functional principal component analysis for longitudinal Riemannian data. Information is pooled across subjects by estimating the mean curve with local Fréchet regression and smoothing the covariance structure of the linearized data on tangent spaces around the mean. Dimension reduction and imputation of the manifold-valued trajectories are achieved by utilizing the leading principal components and applying best linear unbiased prediction. We show that the proposed mean and covariance function estimates achieve state-of-the-art convergence rates. For illustration, we study the development of brain connectivity in a longitudinal cohort of Alzheimer's disease and normal participants by modeling the connectivity on the manifold of symmetric positive definite matrices with the affine-invariant metric. In a second illustration for irregularly recorded longitudinal emotion compositional data for unemployed workers, we show that the proposed method leads to nicely interpretable eigenfunctions and principal component scores. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database.

KEYWORDS

Alzheimer's disease, functional data analysis, longitudinal compositional data, neuroimaging studies, principal component analysis, sampling schemes, symmetric positive-definite matrices

1 | INTRODUCTION

Functional data that assume values in an Euclidean space are typically considered as random elements of a Hilbert space (Horvath and Kokoszka, 2012; Hsing and Eubank, 2015; Wang *et al.*, 2016). Specifically applicable in such linear spaces with their flat Euclidean geometry are key techniques such as functional principal component analysis (Chen and Lei, 2015) and functional regression (Kong *et al.*, 2016). However, only recently has the analysis of nonlinear functional data been considered. In one strand of previous work, the entire functional trajectory is assumed to reside on a nonlinear manifold (Chen and Müller, 2012), while in the other, the response values of the functional or longitudinal data reside on a manifold (Yuan *et al.*, 2012; Lin *et al.*, 2017). The latter setting where data objects are sparsely or densely observed random Riemannian manifold-valued functions is increasingly encountered in practice. Examples where such data are encountered include

neuroimaging (Cornea *et al.*, 2017) and human kinetics studies (Telschow *et al.*, 2019).

Our work is motivated by data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), where one aims to characterize and assess the progression of Alzheimer's disease through collecting longitudinal neuroimaging measures, such as functional magnetic resonance imaging (fMRI) scans. Functional connectivity obtained from resting-state fMRI (rs-fMRI) is known to be altered for Alzheimer's patients (Badhwar *et al.*, 2017) and can be represented by the correlation matrix of brain region activation, which we model on the constrained manifold of symmetric positive-definite matrices. In spite of the relevance for many applied studies of functional connectivity (Ginestet *et al.*, 2017), there is no firm statistical foundation to date for the study of the time-dynamics of data such as longitudinally observed correlation matrices. A second motivating example concerns longitudinal mood assessment of unemployed workers, where in each longitudinal survey, the participants reported the fraction of time during which they experienced four different moods ranging from bad, low, mild, to good. Such longitudinal compositional data consisting of repeated longitudinal observations of compositions that add up to one are encountered in many important applications, eg, repeated voting, consumer preference tracing, soil or air composition over time, and microbiome dynamics.

It has been assumed in virtually all previous studies of manifold-valued functional data (eg, Anirudh *et al.*, 2017; Dai and Müller, 2018; Lin and Yao, 2019) that we are aware of that the recorded data consist of densely or continuously observed functions. A typical example for this are time-varying network data where networks with a fixed number of nodes can be characterized by their graph Laplacians, which are symmetric positive definite matrices, and with this representation networks can be viewed as time-varying metric space-valued objects, a connection that has been made in Dubey and Müller (2020). This connection implies that the manifold methodology of Dai and Müller (2018) and Lin and Yao (2019) is directly applicable for fully observed network trajectories. The Riemannian Functional Principal Analysis through Conditional Expectation (RPACE) approach introduced here then provides a tool to extend these previous results to longitudinal networks, which are irregularly and sparsely observed in time. In general, for most longitudinal biomedical and social studies, data are not continuously observed in time, and more often than not the time points at which measurements are available are irregular and sparse. Nevertheless, one still may assume that the data for each subject result from the realization of an underlying unobserved continuous-time stochastic process that takes values on a Riemannian manifold.

The current lack of suitable methodology for longitudinal manifold-valued data thus provides strong motivation for the development of suitable methods, including a version of principal component analysis, for sparsely observed and potentially noise-contaminated Riemannian longitudinal data. The principal components can subsequently be used to obtain best fits for subject-specific trajectories. Methodology and theory development that is necessary to achieve this goal is challenging, due to the combination of irregularity and scarcity of the repeated measurements that are available per subject on the one hand and the inherent nonlinearity of manifold-valued data on the other hand. Specifically, this development is a highly nontrivial extension of the methodology that is available for longitudinal Euclidean functional data (Yao *et al.*, 2005; Li and Hsing, 2010; Zhang and Wang, 2016).

Functional principal component analysis has emerged as an important tool for analyzing infinite-dimensional functional data. For densely observed manifold-valued functional data, Dai and Müller (2018) studied an intrinsic Riemannian functional principal component analysis that utilizes the Fréchet mean curve defined at each observation time and Riemannian logarithm maps, mapping the observed manifolds at each fixed time to a tangent plane centered around the Fréchet mean of the observed manifolds at that time, and demonstrated its advantages over extrinsic approaches that ignore the manifold structure. While this approach utilized the ambient space in which the manifold is embedded, Lin and Yao (2019) subsequently extended the theory and developed a fully intrinsic approach. In such intrinsic approaches, Fréchet means (Fréchet, 1948) are used to extend the classical Euclidean mean to data on Riemannian manifolds (Patrangenaru *et al.*, 2018).

For sparsely observed manifold-valued longitudinal data, the Riemannian functional principal components approach (Dai and Müller, 2018) that was specifically designed for dense observations is inapplicable. In this work, we instead propose a Riemannian version of the PACE approach (Yao *et al.*, 2005), which aims to pool observations and borrow information across all subjects. For this, we obtain the Fréchet mean curve for longitudinal Riemannian data under a local Fréchet regression framework (Petersen and Müller, 2019), extending both the local linear smoothing paradigm for Euclidean data (Fan and Gijbels, 1996) and for independent Riemannian data (Yuan *et al.*, 2012).

The main innovation and contributions of this paper are as follows: First, we develop a functional principal component analysis for sparsely observed Riemannian longitudinal data. The proposed methods can also handle densely observed Riemannian functional data contaminated with measurement errors, which existing methodology cannot.

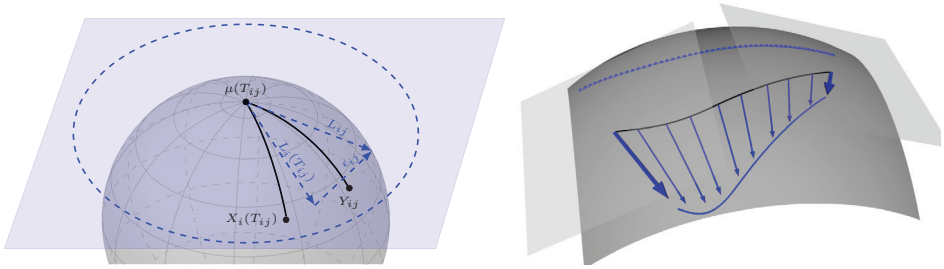

FIGURE 1 Left: Illustration of the tangent space at $\mu(T_{ij})$, $X_i(T_{ij})$, the value of a process on the manifold $\mathcal{M}$ at observation time T_{ij}, the process $L_i(T_{ij})$ that is obtained after applying the logarithm map and lies on the tangent space, and the actually observed value Y_{ij} with noise and its version L_{ij} on the tangent space after applying the logarithm map, illustrating the observation model (5). Specifically, $L_i(T_{ij}) = \mathrm{Exp}_{\mu(T_{ij})} X_i(T_{ij}) \in T_{\mu(T_{ij})}\mathcal{M}$, $L_{ij} = L_i(T_{ij}) + \epsilon_{ij} \in T_{\mu(T_{ij})}\mathcal{M}$ for some random noise $\epsilon_{ij} \in T_{\mu(T_{ij})}\mathcal{M}$, and $Y_{ij} = \mathrm{Exp}_{\mu(T_{ij})} L_{ij} \in \mathcal{M}$. The dashed ellipse in the tangent space $T_{\mu(T_{ij})}\mathcal{M}$ represents the boundary of the domain $\mathcal{D}(\mu(T_{ij}))$. Right: At each time $t \in \mathcal{T}$, the value of the eigenfunction $\phi_k(t)$, denoted by arrows, lies on the tangent space $T_{\mu(t)}$ at the mean $\mu(t)$, a point on the central black curve. The values of the eigenfunction at the first and last time points are bolded and are shown to lie on different tangent spaces. The solid and dotted blue are curves lying on the manifold that represent the mode of variation of μ along the direction of ϕ_k, the kth eigenfunction, which are defined by $\mathrm{Exp}_{\mu(t)}(\phi_k(t))$ and $\mathrm{Exp}_{\mu(t)}(-\phi_k(t))$, $t \in \mathcal{T}$, respectively. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

Second, a local polynomial estimate for the mean function Riemannian longitudinal data is proposed, based on pooling the data across subjects and taking into account the dependency of the repeated measurements within subjects. Third, uniform rates of convergence are derived for the mean and covariance functions for sparse and dense Riemannian functional data under a unified framework, extending previous results for the Euclidean case (Li and Hsing, 2010; Zhang and Wang, 2016) by adopting an M-estimation framework. Fourth, we demonstrate the utility of our methods for longitudinal neuroimaging and social sciences data, as well as in simulation results that are included in the Supporting Information. Finally, an R implementation `RFPCA` is available on GitHub.

2 | METHODOLOGY

2.1 | Preliminaries for manifolds

Let $\mathcal{M}$ be a d-dimensional smooth, connected, and geodesically complete Riemannian manifold isometrically embedded in an ambient space $\mathbb{R}^D$, where positive integers $d \le D$ are the intrinsic and ambient dimensions, respectively. The tangent space $T_p\mathcal{M}$ at $p \in \mathcal{M}$ is a d-dimensional vector space consisting of all velocity vectors $\alpha'(0)$ where $\alpha : (-\epsilon, \epsilon) \to \mathcal{M} \subset \mathbb{R}^D$ is a differentiable curve with $\alpha(0) = p$ defined in a vicinity of 0. The tangent space is endowed with the Riemannian metric $\langle \cdot, \cdot \rangle_p$ induced from the inner product in the ambient Euclidean space, defined by $\langle u, v \rangle_p = \langle u, v \rangle$ for $u, v \in T_p\mathcal{M}$. The geodesic distance $d_{\mathcal{M}}(p, q)$ between $p, q \in \mathcal{M}$ is the infimum of the length taken over all piecewise differentiable curves on $\mathcal{M}$ joining p and q. A geodesic $\gamma : [a, b] \to \mathbb{R}$ is a constant-speed curve on the manifold that is characterized by having a vanishing projection of $\gamma''(t)$ onto $T_{\gamma(t)}$, for $t \in [a, b]$.

The Riemannian exponential map $\mathrm{Exp}_p : T_p\mathcal{M} \to \mathcal{M}$ is defined as $\mathrm{Exp}_p v = \gamma_v(1)$, where $\gamma_v : [0, 1] \to \mathcal{M}$ is a geodesic with initial velocity $\gamma_v'(0) = v$, and the Riemannian logarithm map Log_p is the inverse of Exp_p, assuming that it is well defined; see the left panel of Figure 1. We view $\mathcal{M}$ as a smooth submanifold of the ambient Euclidean space $\mathbb{R}^D$. Since an isometric embedding always exists for a suitable D due to the Nash embedding theorem (Nash, 1956), one does not need to specify an ambient space (Lee, 1997).

2.2 | Statistical model

We define $\mathcal{M}$-valued Riemannian random processes $X(t)$ as a D-dimensional vector-valued random process indexed by a compact interval $\mathcal{T} \subset \mathbb{R}$ such that $X(t) \in \mathcal{M}$, and assume that the process X is of second-order, ie, for every $t \in \mathcal{T}$, there exists $p \in \mathcal{M}$ such that the Fréchet functional $M(p, t) := E[d^2_{\mathcal{M}}(p, X(t))]$ is finite, where $d_{\mathcal{M}}$ is the above-defined geodesic distance. Defining the Fréchet mean function $\mu : \mathcal{T} \to \mathcal{M}$ as

$$\mu(t) = \underset{p \in \mathcal{M}}{\arg\min}\, M(p, t),$$
$$M(p, t) = E\left[d^2_{\mathcal{M}}(p, X(t))\right], \quad t \in \mathcal{T}, \tag{1}$$

the following assumption is required to ensure the existence and uniqueness of the Fréchet mean curve:

(X0) X is of second-order, and the Fréchet mean curve $\mu(t)$ exists and is unique.

Assumption (X0) holds, for example, if the range of the process X is contained in a geodesic ball of sufficient small radius (Afsari, 2011). Since it is assumed that $\mathcal{M}$ is geodesically complete, by the Hopf—Rinow theorem, its exponential map Exp_p at each p is defined on the entire tangent space $T_p\mathcal{M}$ at $p \in \mathcal{M}$. We define the domain $\mathcal{D}_p$ to be the interior of the collection of tangent vectors $v \in T_p\mathcal{M}$ such that if $\gamma(t) = \mathrm{Exp}_p tv$ is a geodesic emanating from p with the direction v, then $\gamma([0,1])$ is a minimizing geodesic. On the domain $\mathcal{D}_p$, the map Exp_p is injective with image $\mathrm{Im}(\mathrm{Exp}_p)$. The Riemannian logarithm map at p, denoted by Log_p, is the inverse of Exp_p, restricted to $\mathrm{Im}(\mathrm{Exp}_p)$; if $q = \mathrm{Exp}_p v$ for some $v \in \mathcal{D}_p$, then $\mathrm{Log}_p q = v$. The logarithm map Log_p effectively provides a local linear approximation of a neighborhood of $p \in \mathcal{M}$, mapping on the tangent $T_p\mathcal{M}$.

To study the covariance structure of the random process X on tangent spaces, we require

(X1) For some constant $\varepsilon_0 > 0$, $X(t) \in \mathcal{M}\backslash\{\mathcal{M}\backslash\mathrm{Im}(\mathrm{Exp}_{\mu(t)})\}^{\varepsilon_0}$ for $t \in \mathcal{T}$, where A^{ε_0} denotes the set $\bigcup_{p\in A}\{q \in \mathcal{M} : d_{\mathcal{M}}(p,q) < \varepsilon_0\}$.

This condition requires $X(t)$ to stay away from the cut locus (Lee, 1997) of $\mu(t)$ uniformly for $t \in \mathcal{T}$, so that the logarithm map $\mathrm{Log}_{\mu(t)}$ is well defined for all t. It is not needed if $\mathrm{Exp}_{\mu(t)}$ is injective on $T_{\mu(t)}\mathcal{M}$ for all t. In the special case of a d-dimensional unit sphere $\mathbb{S}^d$ that we deal with for the special case of longitudinal compositional data, if $X(t)$ is continuous and the distribution of $X(t)$ vanishes at an open set with positive volume that contains $\mathcal{M}\backslash\mathrm{Im}(\mathrm{Exp}_{\mu(t)})$, (X1) holds. Under (X0) and (X1), the $\mathbb{R}^D$-valued logarithm process $L(t) := \mathrm{Log}_{\mu(t)}X(t)$ is well defined for all $t \in \mathcal{T}$.

An important observation (Bhattacharya and Patrangenaru, 2003) is that $EL(\cdot) = 0$, and furthermore, that $E\|L(t)\|_2^2 = Ed^2_{\mathcal{M}}(\mu(t), X(t)) < \infty$ for every $t \in \mathcal{T}$, where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^D$, which makes it possible to define the covariance function of L by

$$\Gamma(s,t) = E[L(s)L(t)^{\mathrm{T}}], \qquad s,t \in \mathcal{T}. \tag{2}$$

This covariance function and its covariance operator admit the eigendecomposition

$$\Gamma(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k^{\mathrm{T}}(t)$$

with orthonormal eigenfunctions ϕ_k and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, where $\sum_{k=1}^{\infty}\lambda_k < \infty$, leading to the Karhunen—Loève representation (Grenander, 1950; Kleffe, 1973)

$$L(t) = \sum_{k=1}^{\infty} \xi_k \phi_k(t), \quad \xi_k = \int_{\mathcal{T}} L(t)^{\mathrm{T}}\phi_k(t)dt, \tag{3}$$

where the ξ_k are the uncorrelated Riemannian functional principal component scores with $E\xi_k = 0$ and $E\xi_k^2 = \lambda_k$. We will utilize finitely truncated versions

$$X_K(t) := \mathrm{Exp}_{\mu(t)}L_K(t), \quad L_K(t) = \sum_{k=1}^{K}\xi_k\phi_k(t) \tag{4}$$

for a given integer $K \geq 0$. The eigenfunctions ϕ_k, defined on the tangent spaces, represent the linearized modes of variation of X; see the right panel of Figure 1.

To reflect longitudinal studies, for a sample $X_1, \ldots, X_n$ of an $\mathcal{M}$-valued Riemannian random process X, we assume that each X_i is only recorded at m_i random time points $T_{i1}, \ldots, T_{im_i} \in \mathcal{T}$, where each observation $X_i(T_{ij})$ is additionally corrupted by intrinsic random noise. Specifically, with $\mathbb{D}_n = \{(T_{ij}, Y_{ij}) : i = 1, \ldots, n;\ j = 1, \ldots, m_i\}$, one records noisy measurements

$$Y_{ij} = \mathrm{Exp}_{\mu(T_{ij})}L_{ij}, \quad L_{ij} = L_i(T_{ij}) + \epsilon_{ij}, \tag{5}$$

where measurement times T_{ij} are identically distributed and independent of the predictors X_i, with $T_{ij} \sim f$ for some density f supported on $\mathcal{T}$. A graphical illustration of the sampling model (5) is in the left panel of Figure 1. We require

(Y0) Conditional on $\{T_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, m_i\}$, the ϵ_{ij} are independent and are independent of the X_i, with isotropic variance σ^2 and $E(\epsilon_{ij} \mid T_{ij}) \equiv 0$. Furthermore, the Fréchet mean of Y_{ij} conditional on T_{ij} is $\mu(T_{ij})$.

As $E(L_i(T_{ij}) \mid T_{ij}) = 0$, the assumption on ϵ_{ij} implies that $E(L_{ij} \mid T_{ij}) = 0$. The random noise ϵ_{ij}, although modeled in the tangent spaces, induces random noise on the manifold itself via Riemannian exponential maps, and could alternatively be modeled directly on the manifold, under the centering condition that the Fréchet mean of Y_{ij} given T_{ij} is $\mu(T_{ij})$, which is the equivalent of the usual centering condition for model errors in Euclidean space. The following condition is analogous to (X1) and is needed for Y_{ij} to stay away from the cut locus. It imposes indirect constraints on the random noise. For example, it requires the random noise to be bounded when the cut locus is

not empty. The condition is not needed when the underlying manifold $\mathcal{M}$ has nonpositive sectional curvature (Lee, 1997).

(Y1) For some constant $\varepsilon_1 > 0$, $Y_{ij} \in \mathcal{M} \backslash \{\mathcal{M} \backslash \mathrm{Im}(\mathrm{Exp}_{\mu(T_{ij})})\}^{\varepsilon_1}$ for $i = 1, \ldots, n; j = 1, \ldots, m_i$.

2.3 | Estimation

To deal with the sparse and irregularly spaced observations coming from a longitudinal study, we develop a new method for estimating the Fréchet mean function by harnessing the Fréchet regression framework that was originally designed for the case of independent observations (Petersen and Müller, 2019). We study an extension that is valid for dependent repeated measurements in a unified framework that covers both sparse and dense observations.

To construct a local polynomial smoother for manifold-valued responses, we define the local weight function at $t \in \mathcal{T}$ as

$$\hat{\omega}_{ij}(t, h_\mu) = \frac{1}{\hat{\sigma}_0^2} K_{h_\mu}(T_{ij} - t)\{\hat{u}_2 - \hat{u}_1(T_{ij} - t)\},$$

where $\hat{\sigma}_0^2(t) = \hat{u}_0(t)\hat{u}_2(t) - \hat{u}_1^2(t)$, $\hat{u}_k(t) = \sum_{i=1}^n w_i \sum_{j=1}^{m_i} K_{h_\mu}(T_{ij} - t)(T_{ij} - t)^k$ for $k = 0, 1, 2$, w_i are subject-specific weights satisfying $\sum_{i=1}^n m_i w_i = 1$, $K_{h_\mu}(\cdot) = h_\mu^{-1} K(\cdot/h_\mu)$, $K(\cdot)$ is a symmetric density function, and $h_\mu > 0$ is a bandwidth. The mean $\mu(t)$ is estimated by

$$\hat{\mu}(t) = \underset{y \in \mathcal{M}}{\arg\min}\ Q_n(y, t),$$

where we define the double-weighted Fréchet function as

$$Q_n(y, t) = \sum_{i=1}^n w_i \sum_{j=1}^{m_i} \hat{\omega}_{ij}(t, h) d_{\mathcal{M}}^2(Y_{ij}, y),$$

for $y \in \mathcal{M}$ and $t \in \mathcal{T}$. For the special case $\mathcal{M} = \mathbb{R}^D$ where observations lie in a Euclidean space, Q_n coincides with the sum of squared errors loss used in Zhang and Wang (2016). Two prominent schemes are to assign equal weight to each observation, resulting in $w_i = (n\bar{m})^{-1}$, $\bar{m} = n^{-1}\sum_{i=1}^n m_i$ (Yao *et al.*, 2005) or to assign equal weight to each subject, ie, $w_i = (nm_i)^{-1}$ (Li and Hsing, 2010). As for the Euclidean case, the former scheme was found to work better for non-dense and the latter for ultradense functional data, and hence we will adopt the former weight scheme in our implementations for sparsely sampled data.

To estimate the covariance structure, we first map the observed data into tangent spaces, obtaining D-dimensional column vectors $\hat{U}_{ij} = \mathrm{Log}_{\hat{\mu}(T_{ij})} Y_{ij}$ and then the matrix-valued covariance function Γ by scatterplot smoothing (Yao *et al.*, 2005) for matrix-valued data matrices $\Gamma_{ijl} = \hat{U}_{ij}\hat{U}_{il}^{\mathrm{T}}$ for $j \neq l$. This leads to the estimates $\hat{\Gamma}(s, t) = \hat{A}_0$, where

$$\begin{aligned}(\hat{A}_0, \hat{A}_1, \hat{A}_2) &:= \underset{A_0, A_1, A_2}{\arg\min} \sum_{i=1}^n \upsilon_i \sum_{1 \le j \neq l \le m_i} K_{h_\Gamma}(T_{ij} - s) K_{h_\Gamma}(T_{il} - t) \\ &\quad \times \|\Gamma_{ijl} - A_0 - (T_{ij} - s)A_1 - (T_{il} - t)A_2\|_F^2. \end{aligned} \tag{6}$$

Here, $\|\cdot\|_F$ is the matrix Frobenius norm, $h_\Gamma > 0$ is a bandwidth, and υ_i are weights satisfying $\sum_{i=1}^n m_i(m_i - 1)\upsilon_i = 1$, where we select $\upsilon_i = 1/\sum_{i=1}^n m_i(m_i - 1)$ in the sparse longitudinal case; see Zhang and Wang (2016, 2018) for other possible choices. Estimates for the eigenfunctions ϕ_k and λ_k of Γ are then obtained by the eigenfunctions $\hat{\phi}_k$ and eigenvalues $\hat{\lambda}_k$ of $\hat{\Gamma}$.

In applications, one needs to choose appropriate bandwidths h_μ and h_Γ, as well as the number of included components K. To select h_μ for smoothing the mean function μ, we adopt a generalized cross-validation (GCV) criterion

$$\mathrm{GCV}(h) = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} d_{\mathcal{M}}^2(\hat{\mu}(T_{ij}), Y_{ij})}{(1 - K_h(0)/N)^2},$$

where $N = \sum_{i=1}^n m_i$ is the total number of observations, selecting h_μ as the minimizer of $\mathrm{GCV}(h)$. While a similar GCV strategy can be applied to select the bandwidth h_Γ for the covariance function, we propose to employ the simpler choice $h_\Gamma = 2h_\mu$, which we found to be computationally efficient and to perform well in simulations. To practically determine the number of components K included in the finite-truncated representation (4), we consider the fraction of variation explained (FVE)

$$\mathrm{FVE}(K) = \frac{\sum_{k=1}^K \lambda_k}{\sum_{j=1}^\infty \lambda_j}, \quad \widehat{\mathrm{FVE}}(K) = \frac{\sum_{k=1}^K \hat{\lambda}_k}{\sum_{j=1}^\infty \hat{\lambda}_j}, \tag{7}$$

and choose the number of included components as the smallest K such that FVE exceeds a specified threshold $0 < \gamma < 1$,

$$\begin{aligned} K^* &= \min(K : \mathrm{FVE}(K) \ge \gamma), \\ \hat{K}^* &= \min(K : \widehat{\mathrm{FVE}}(K) \ge \gamma). \end{aligned} \tag{8}$$

Commonly γ is set to .90, .95, or .99.

2.4 | Riemannian functional principal component analysis through conditional expectation

The unobserved Riemannian functional principal component scores ξ_{ik} (4) need to be estimated from the discrete samples $\{(T_{ij}, Y_{ij})\}_{j=1}^{m_i}$. Approximating the integral in (3) is infeasible when the number of repeated measurements per curve is small, which is well known for the Euclidean case (Yao *et al.*, 2005). We therefore develop in the following RPACE for tangent vector-valued processes. Throughout this subsection, expected values will be taken conditional on the observation time points $\{T_{ij}\}_{i=1\ j=1}^{n\ m_i}$.

Applying best linear unbiased prediction of ξ_{ik}, we obtain scores

$$\tilde{\xi}_{ik} = \mathcal{B}(\xi_{ik} \mid U_i) = \lambda_k \phi_{ik}^{\mathrm{T}} \Sigma_{U_i}^{-1} U_i, \tag{9}$$

where $\mathcal{B}$ denotes the best linear unbiased predictor; with the vectorization operation denoted as $\mathrm{Vec}(\cdot)$, $U_i = \mathrm{Vec}((U_{i1}, \ldots, U_{im_i}))$ are the vectorized log-mapped noisy observations for subject i, $\tilde{U}_i = \mathrm{Vec}[(L_i(T_{i1}), \ldots, L_i(T_{im_i}))]$, $\phi_{ik} = \mathrm{Vec}[(\phi_k(T_{i1}), \ldots, \phi_k(T_{im_i}))]$, and $\Sigma_{U_i} = E(U_i U_i^{\mathrm{T}}) = E(\tilde{U}_i \tilde{U}_i^{\mathrm{T}}) + \sigma^2 I$, where I is the identity matrix. The entry of $E(\tilde{U}_i \tilde{U}_i^{\mathrm{T}})$ corresponding to $E[\{L_i(T_{ij})\}_l \{L_i(T_{il})\}_m]$ is $\{\Gamma(T_{ij}, T_{ik})\}_{lm}$, where $\{v\}_a$ and $\{A\}_{ab}$ denote the ath and (a,b)th entry in a vector v and matrix A, respectively. Estimates $\tilde{\xi}_{ik}$ coincide with the conditional expectations $E(\xi_{ik} \mid U_i)$ if the joint distribution of (ξ_{ik}, U_i) is elliptically contoured (Fang *et al.*, 1990, Theorem 2.18) such as the Gaussian distribution.

Substituting estimates for the corresponding unknown quantities in (9), we obtain

$$\hat{\tilde{\xi}}_{ik} = \hat{\lambda}_k \hat{\phi}_{ik}^{\mathrm{T}} \hat{\Sigma}_{U_i}^{-1} \hat{U}_i, \tag{10}$$

where $\hat{\Sigma}_{U_i} = \hat{E}(\tilde{U}_i \tilde{U}_i^{\mathrm{T}}) + \hat{\sigma}^2 I$ and $\hat{E}(\tilde{U}_i \tilde{U}_i^{\mathrm{T}})$, $\hat{\lambda}_k$ and $\hat{\phi}_{ik}$ are obtained from $\hat{\Gamma}$, the minimizer of (6), and $\hat{\sigma}^2 = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (ndm_i)^{-1} \mathrm{tr}[L_{ij} L_{ij}^{\mathrm{T}} - \hat{\Gamma}(T_{ij}, T_{ij})]$. The K-truncated processes

$$L_{iK}(t) = \sum_{k=1}^{K} \xi_{ik} \phi_k(t), \quad X_{iK}(t) = \sum_{k=1}^{K} \mathrm{Exp}_{\mu(t)} L_{iK}(t) \tag{11}$$

are then estimated by

$$\hat{L}_{iK}(t) = \sum_{k=1}^{K} \hat{\tilde{\xi}}_{ik} \hat{\phi}_k(t), \quad \hat{X}_{iK}(t) = \sum_{k=1}^{K} \mathrm{Exp}_{\hat{\mu}(t)} \hat{L}_{iK}(t). \tag{12}$$

If one aims to estimate the underlying processes $X_i(t)$ or $L_i(t)$ at some fixed $t \in \mathcal{T}$, one can also proceed directly without estimating the scores. Best linear unbiased predictors are obtained as

$$\tilde{L}_i(t) = \mathcal{B}(L_i(t) \mid U_i) = \Sigma_{it}^{\mathrm{T}} \Sigma_{U_i}^{-1} U_i, \quad \tilde{X}_i(t) = \mathrm{Exp}_{\mu(t)} \tilde{L}_i(t), \tag{13}$$

with corresponding estimates

$$\hat{L}_i(t) = \hat{\Sigma}_{it}^{\mathrm{T}} \hat{\Sigma}_{U_i}^{-1} \hat{U}_i, \quad \hat{X}_i(t) = \mathrm{Exp}_{\hat{\mu}(t)} \hat{L}_i(t), \tag{14}$$

obliviating the need for finite truncation. Here, $\Sigma_{it} = E[U_i L_i(t)^{\mathrm{T}}] = (\Gamma(T_{i1}, t)^{\mathrm{T}}, \ldots, \Gamma(T_{im_i}, t)^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\Sigma}_{it}$ is the plug-in estimate of Σ_{it}.

3 | ASYMPTOTIC PROPERTIES

To derive the asymptotic properties of the estimates in Section 2, we require the following assumptions, in addition to conditions (X0) and (X1); see the Supporting Information for details.

(M0) The domain $\mathcal{T}$ is a compact interval and $\mathcal{M}$ is a bounded submanifold of $\mathbb{R}^D$.

(K0) The kernel function K is a Lipschitz continuous symmetric probability density function on $[-1, 1]$.

(X2) Almost surely, the sample paths $X(\cdot)$ are twice continuously differentiable.

(X3) The density f of the T_{ij} is positive and twice continuously differentiable.

Boundedness of the manifold as in (M0) can be replaced by compact support conditions on the random process $X(t), t \in \mathcal{T}$ and the noisy observations Y_{ij}. Conditions (X2) and (X3) concern the smoothness of the process and design density and are standard for the Euclidean case (Zhang and Wang, 2016).

Let $\omega(s,t,h) = \sigma_0(t)^{-2} K_h(s-t)\{u_2(t) - u_1(t)(s-t)\}$, where $u_k(t) = E\{K_h(T-t)(T-t)^k\}$, $k = 0, 1, 2$, and $\sigma_0^2(t) = u_0(t)u_2(t) - u_1^2(t) > 0$, $t \in \mathcal{T}$ by the Cauchy—Schwarz inequality. The finiteness of u_k is implied by the Lipschitz continuity of the kernel function K and the compactness of the domain $\mathcal{T}$. Define $\tilde{Q}_h(p,t) = E(\omega(T,t,h) d_{\mathcal{M}}^2(Y,p))$ and $\tilde{\mu}(t) = \arg\min_{y \in \mathcal{M}} \tilde{Q}_h(y,t)$. Two additional conditions (L0) and (L1) are needed.

(L0) The Fréchet mean functions $\tilde{\mu}$ and $\hat{\mu}$ exist and are unique, the latter almost surely for all n.

Defining a real-valued function $G_p(v,t) = \tilde{M}(\mathrm{Exp}_p v, t)$, $v \in T_p\mathcal{M}$ and $t \in \mathcal{T}$, where $\tilde{M}(p,t) = E(d^2_{\mathcal{M}}(p, Y_{ij}) \mid T_{ij} = t)$ for $p \in \mathcal{M}$, $T_p\mathcal{M}$ denotes as before the tangent space at p and $\mathrm{Exp}_p : T_p\mathcal{M} \to \mathcal{M}$ the Riemannian exponential map, we assume

(L1) The Hessian of $G_p(\cdot, t)$ at $v = 0$ is uniformly positive definite along the mean function, ie, for its smallest eigenvalue $\lambda_{\min}$ it holds that

$$\inf_{t\in\mathcal{T}} \lambda_{\min}\left\{ \frac{\partial^2}{\partial v^2} G_{\mu(t)}(v,t) \mid_{v=0} \right\} > 0.$$

Conditions (L0) and (L1) ensure properly defined minima and are necessary for consistent estimation of the mean curve using M-estimation theory (Petersen and Müller, 2019). On a Riemannian manifold $\mathcal{M}$ with sectional curvature at most $\mathcal{K}$, (L0) and (L1) are satisfied asymptotically if the support of the noisy observations Y_{ij} in the local time window stays within $B\{\mu(t), \pi/(2\mathcal{K})\}$, where $B(p,r)$ is a geodesic ball with center $p \subset \mathcal{M}$ and radius r (Bhattacharya and Bhattacharya, 2012); this specifically holds for longitudinal compositional data mapped to the positive orthant of a unit sphere. The next two conditions on the bandwidths h_μ and h_Γ are needed to derive the rate of convergence of the mean and covariance estimates, respectively. For simplicity of presentation, we assume $m_i = m$, noting that extensions to more general cases are straightforward (Zhang and Wang, 2016).

(H1) $h_\mu \to 0$ and $(\log n)/(nmh_\mu) \to 0$.
(H2) $h_\Gamma \to 0$, $(\log n)/(nm^2h_\Gamma^2) \to 0$, and $(\log n)/(nmh_\Gamma) \to 0$.

Theorem 1. *Assume that conditions (X0)-(X3), (Y0)-(Y1), (M0), (K0), (L0)-(L1), and (H1) hold. Then,*

$$\sup_{t\in\mathcal{T}} d^2_{\mathcal{M}}\{\hat{\mu}(t), \mu(t)\} = O_P\left(h_\mu^4 + \frac{\log n}{nmh_\mu} + \frac{\log n}{n} \right). \quad (15)$$

Theorem 1 shows that estimate $\hat{\mu}$ enjoys the same uniform convergence rate as in Zhang and Wang (2016) for the Euclidean case, so the presence of curvature does not impact the rate. The rate in (15) has three terms that correspond to three regimes that are characterized by the growth rate of m relative to the sample size: (a) When $m \ll (n/\log n)^{1/4}$, the observations per curve are sparse and the optimal choice $h_\mu \asymp (\log n/nm)^{1/5}$ yields $\sup_{t\in\mathcal{T}} d_{\mathcal{M}}\{\hat{\mu}(t), \mu(t)\} = O_P\{(\log n/nm)^{2/5}\}$. (b) When $m \asymp (n/\log n)^{1/4}$, corresponding to an intermediate case, the optimal choice $h_\mu \asymp (\log n/n)^{1/4}$ leads to the uniform rate $O_P\{(\log n/n)^{1/2}\}$ for $\hat{\mu}$. (c) When $m \gg (n/\log n)^{1/4}$, the observations are dense, and any choice $h_\mu = o\{(\log n/n)^{1/4}\}$ gives rise to the uniform rate $O_P\{(\log n/n)^{1/2}\}$.

We note that the transition from (a) to (c) is akin to a phase transition as observed in Hall *et al.* (2006). Our next result concerns the uniform rate for the estimator $\hat{\Gamma}$ of Γ, the covariance function of the log-mapped data, extending a result of Zhang and Wang (2016) to manifold-valued functional data.

Theorem 2. *Assume conditions (X0)-(X3), (Y0)-(Y1), (M0), (K0), (L0)-(L1), (H1), and (H2) hold. Then,*

$$\sup_{s,t\in\mathcal{T}} \|\hat{\Gamma}(s,t) - \Gamma(s,t)\|_F^2 = O_P\left(h_\mu^4 + h_\Gamma^4 + \frac{\log n}{nmh_\mu} + \frac{\log n}{n} + \frac{\log n}{nm^2h_\Gamma^2} + \frac{\log n}{nmh_\Gamma} \right). \quad (16)$$

Again, the above rate gives rise to three regimes that are determined by the growth rate of m relative to the sample size and are discussed in Web Appendix C in the Supporting Information, along with its implications for estimated eigenvalues and eigenfunctions, where corresponding rates are obtained by applying results from perturbation theory (Bosq, 2000).

4 | DATA APPLICATIONS

4.1 | Time-evolving functional connectivity

The longitudinal development of brain functional connectivity, defined as the temporal dependency of neuronal activation patterns in different brain regions, has become a focus of recent investigations in brain imaging (Deoni *et al.*, 2016; Dai *et al.*, 2019) to quantify brain development and brain aging. In such studies, brain imaging modalities that include fMRI scans are often obtained longitudinally to quantify the coactivation of various brain regions over time, where activation of a region is inferred from elevated blood oxygen levels in the specified region. We focus here on quantifying the effects of brain aging in terms of longitudinally varying functional connectivity between brain regions, assessed using fMRI scans that are obtained from subjects in a relaxed state. Brain connectivity is measured by calculating a correlation (Worsley *et al.*, 2005) that essentially corresponds to functional dynamic correlation (Dubin and Müller, 2005) between the average signals of various brain regions. The observed correlations are

frequently contaminated with measurement errors from various sources (Laumann *et al.*, 2017).

Recent work on longitudinal modeling of fMRI-based correlation matrices and connectivity has focused on linear models (Hart *et al.*, 2018) with their associated restrictions or on graph-based methods under very limited designs where just two repeated observations are considered per subject (Kundu *et al.*, 2019). In contrast to these previous approaches that are grounded in traditional Euclidean representations, we propose here a nonparametric approach for general longitudinal data that is highly flexible and is designed for objects that form Riemannian manifolds and where the fitted trajectories automatically stay in the same space; a directly comparable method does not exist yet. Obtaining trajectories of correlation matrices first, these can then be converted into time-varying dynamic connectivity measures.

The data for our analysis are from ADNI, a longitudinal study that has recorded repeated rs-fMRI scans. Of central interest are changes in brain function for Alzheimer's patients after the onset of the disease. Since the time of onset is unobservable, we chose as a proxy the time of the first scan for each subject, the first time at which the diagnosis status is available, and for all subjects, we chose this approximate onset time as the origin of the time axis. The times T_{ij}, $j = 1, \ldots, m_i$, when fMRI scans were obtained for the ith subject, are accordingly recorded relative to the time origin at $t = 0$, which means that the first scan for each subject takes place at the time $T_{i1} = 0$. The raw rs-fMRI data were first preprocessed by following a standard protocol that involves motion correction, slice timing correction, coregistration, normalization, and detrending. For each scan, we obtained the pairwise correlations between 10 brain regions that were identified as relevant for brain connectivity in Buckner *et al.* (2009). The resulting 10×10 correlation matrices constitute the manifold-valued responses Y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, at each scan time T_{ij}. For each subject, one thus has a random number of correlation matrices that are sparsely observed in time and may be noise-contaminated.

It is of interest to determine and compare the continuously interpolated mean trajectories for both Alzheimer's and normal subjects, for which we apply the proposed RPACE. The correlation matrices Y_{ij} are modeled on the Riemannian manifold of symmetric positive definite matrices $\mathcal{M} = \mathrm{SPD}(10) = \{\Sigma \in \mathbb{R}^{10\times 10} : \Sigma \text{ is symmetric positive definite}\}$, for which the tangent space $T_P\mathcal{M}$ for a $P \in \mathcal{M}$ is represented by the collection of 10×10 symmetric matrices. We employ the affine-invariant Riemannian metric on $\mathcal{M}$ (Pennec *et al.*, 2006), which can be defined through $\langle U, V\rangle = \mathrm{tr}(UV)$ for $U, V \in T_P\mathcal{M}$. While the representation of tangent spaces does not depend on the selected P, the Riemannian metric, exponential maps, and logarithm maps depend on P, as follows. For a symmetric positive definite $P \in \mathcal{M}$, the Riemannian exponential map Exp_P takes a symmetric matrix $V \in T_P\mathcal{M}$ to a symmetric positive definite matrix via $\mathrm{Exp}_P V = P^{1/2}\,\mathrm{expm}(V)P^{1/2}$, and the Riemannian logarithm map Log_P takes a symmetric positive definite $Q \in \mathcal{M}$ to a symmetric matrix on $T_P\mathcal{M}$ via $\mathrm{Log}_P Q = \mathrm{logm}(P^{-1/2}QP^{-1/2})$, where $\mathrm{expm}(\cdot)$ and $\mathrm{logm}(\cdot)$ are the matrix exponential and logarithm. The geodesic distance is $d_{\mathcal{M}}(P, Q) = \|\,\mathrm{logm}(P^{-1/2}QP^{-1/2})\|_F$, where $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$ denotes the Frobenius norm of a matrix $A = (a_{ij})$. The proposed methods guarantee that the fitted objects lie on the $\mathcal{M}$ and are always SPD matrices, which can then be converted to correlation matrices. This is an important feature that distinguishes the proposed geometric approach from classical Euclidean methods, where there is no such guarantee. The affine-invariant metric endows $\mathcal{M}$ with globally nonpositive curvature (Pennec *et al.*, 2006), which ensures that Fréchet mean, exponential map, and logarithm map are always well defined.

The time window of interest in the longitudinal connectivity analysis is $\mathcal{T} = [0, 1.1]$ years after the initial visit at $t = 0$, where the time domain is chosen to allow at least one full year of observations. After removing subjects with outlying signals or no repeated rs-fMRI measurements within 1.1 years of the initial visit, the sample consisted of 64 subjects, of whom 26 had a diagnosis of Alzheimer's disease and 38 were cognitively normal. A total number of 215 scans were available with two to four repeated measurements per subject. The sparsity and irregularity of these data poses difficulties for classical analyses, prevents the application of presmoothing, and renders previous approaches (Dai and Müller, 2018) infeasible. The proposed RPACE method is geared toward such sparse and irregularly sampled manifold-valued functional data and guarantees consistent estimation.

The raw correlation matrices Y_{ij} for three randomly selected subjects are displayed in the first row of each panel in Figure 2. The large heterogeneity in the raw correlations suggests the presence of substantial measurement errors. The eigenvalues decay slowly in this example, motivating the application of (14) to obtain noise-filtered fitted trajectories $\hat{X}_i(t)$ without resorting to finite-dimensional truncation. The fitted correlations with $h_\mu = h_\Gamma = 0.3$ and the Gaussian kernel, displayed in the second row of each panel in Figure 2, are clearly smoother and less noisy compared to the raw correlations, which helps to delineate underlying trends.

To further demonstrate the application of our methodology for comparisons of SPD(10)-valued connectivity matrices, we focus on the initial visit time $t = 0$, where a noisy raw correlation matrix Y_{i1} is available for each subject,

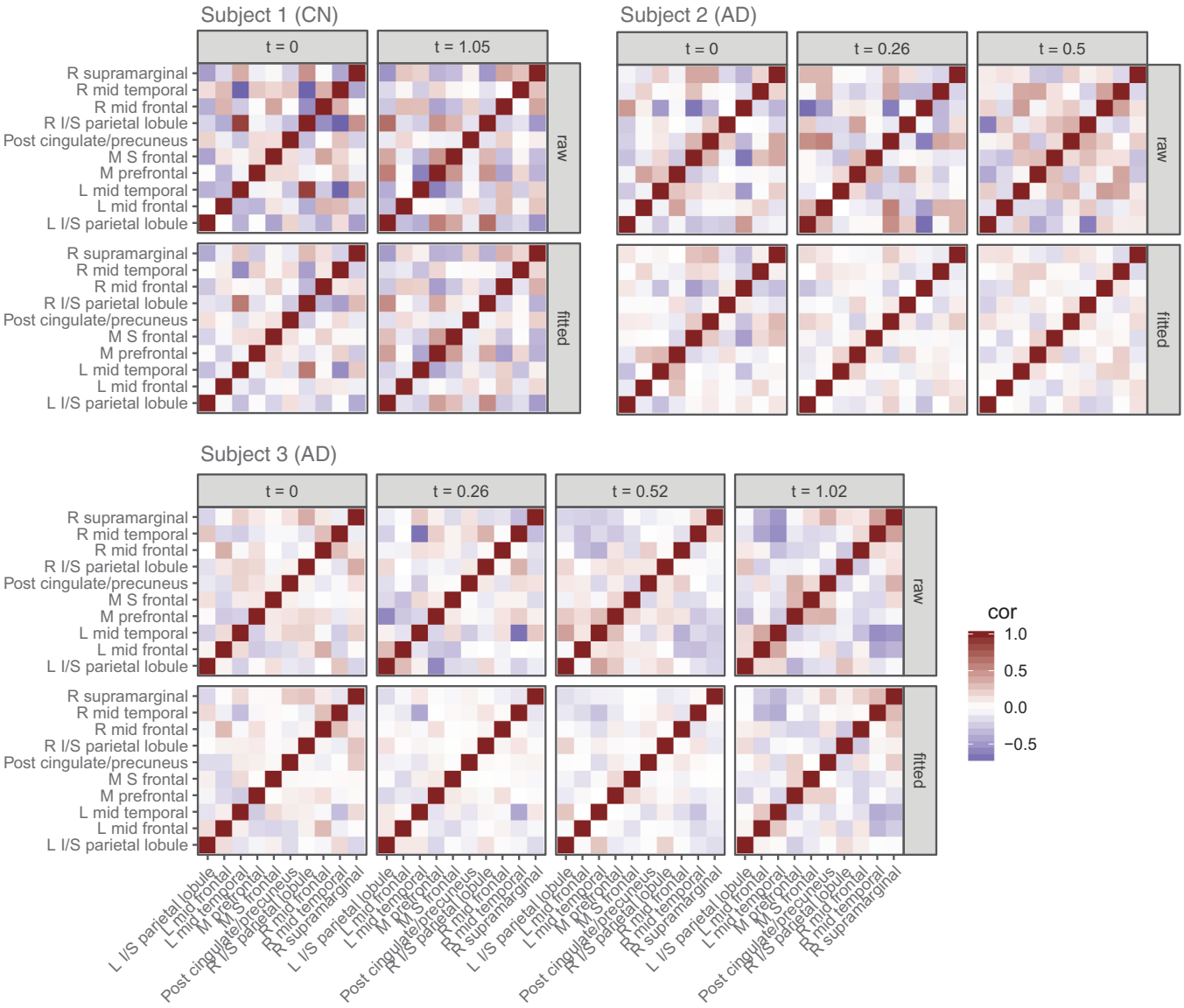

FIGURE 2 Longitudinal functional connectivity of three randomly selected subjects, reflected by 10 × 10 correlation matrices for ten brain regions, where the subject in the left upper panel is cognitively normal and has two available measurements and the subjects in the right upper and lower panels have been diagnosed with Alzheimer's and have three and, respectively, four, available measurements, with times of the measurements in years as indicated in the panels. Each panel includes observed (top row) and fitted (bottom row) correlation matrices. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

as well as the value of the noise-filtered fitted trajectory $\hat{X}_i(0)$. For a simple-minded approach, one could also consider the average connectivity matrix $\bar{X}_i$, defined as the Fréchet mean of all correlation matrices observed at random times for the ith subject. For each of the raw, fitted, and average correlation matrices available per subject, we compare the correlation matrices of the subjects with Alzheimer's with those of the cognitively normal group. We apply the Fréchet analysis of variance test of Dubey and Müller (2019) for the two sample comparison of the correlation matrices, where we refer to this paper for further details about the test. The P-values obtained with the bootstrap version of the test are $P = .03$ when using the fitted correlation matrices $\hat{X}_i(0)$, $P = .37$ when using the raw correlation matrices at the first visit Y_{i1}, and $P = .47$ when using the subject-specific correlation matrix averages $\bar{X}_i$. This suggests that the noise reduction achieved by the proposed fitting procedure leads to more powerful inference.

4.2 | Emotional well-being for unemployed workers

In this second data example, we analyze data from the Survey of Unemployed Workers in New Jersey (Krueger

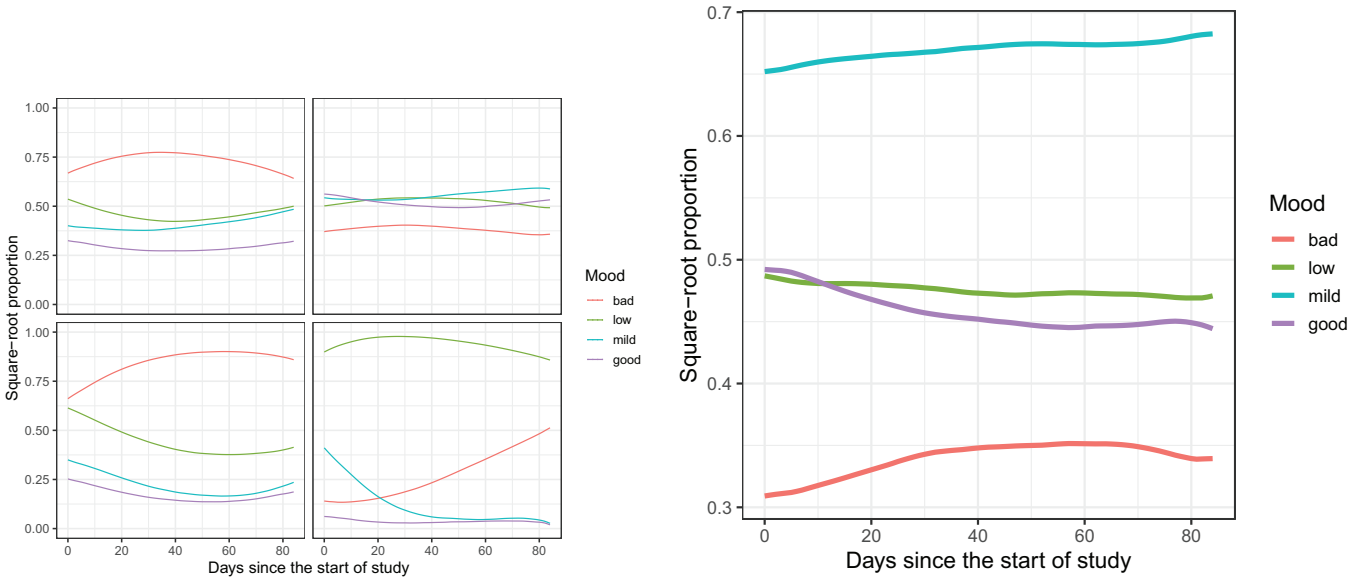

FIGURE 3 Left: Longitudinal mood compositional data for four randomly selected unemployed workers, with raw observations shown as dots and fitted trajectories by the proposed method shown as solid curves, using eight eigencomponents. Overlapping dots were slightly jittered vertically. Right: The overall mean function. All functions are shown in the square root transformation scale. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

and Mueller, 2011) conducted in the fall of 2009 and the beginning of 2010, during which the unemployment rate in the United States peaked at 10% after the 2007-2008 financial crisis. The data are from a stratified random sample of unemployed workers, who were surveyed weekly for up to 12 weeks. Questionnaires included an entry survey, which assessed baseline characteristics such as household income, and weekly follow-ups regarding job search activities and emotional well-being. In each follow-up questionnaire, participants were asked to report the percentage of time they spent in each of four different moods.

We consider a sample of $n = 4771$ unemployed workers enrolled in the study who were not offered a job during the survey period. Times T_{ij} at which subject i responded to the jth survey were recorded as days since the start of the study that falls within $\mathcal{T} = [0, 84]$. The overall weekly response rate was around 40% and the number of responses m_i per subject ranged from 1 to 12, with 25% of all subjects having only one response recorded. Thus these data are a mixture of very sparse and somewhat sparse longitudinal observations. As subjects responded on different days of the week, the times T_{ij} at which the compositional mood vector was recorded are not only sparse but also irregularly spaced. At each T_{ij}, compositional data $Z_{ij} = (Z_{ij1}, \dots, Z_{ij4})$, $j = 1, \dots, m_i$, are observed, where Z_{ijl} is the reported and assumed to be noise-contaminated proportion of time subject i spent in the lth mood in the previous week, $l = 1, \dots, 4$ corresponding to bad, low/irritable, mildly pleasant, and good moods. The Z_{ij} reflect an underlying mood composition process $W_i(t) = \{W_{i1}(t), \dots, W_{i4}(t)\}$, where $W_{il}(t)$ is the proportion of time a subject spent in the lth mood in the week preceding day t.

The proposed RPACE method was applied to the square-root transformed compositional data Y_{ij} and compositional process X_i, defined as

$$Y_{ij} = \left(\sqrt{Z_{ij1}}, \dots, \sqrt{Z_{ij4}}\right),$$

$$X_i(t) = \left(\sqrt{W_{i1}(t)}, \dots, \sqrt{W_{i4}(t)}\right),$$

both lying on the sphere $\mathbb{S}^3$ for $t \in [0, 84]$, as compositional data are nonnegative and sum to one (Scealy and Welsh, 2011; Dai and Müller, 2018). Two geometries might be considered as alternatives to the proposed spherical geometry: the Aitchison geometry (Aitchison, 1986) obtained through applying a log-ratio transformation and the Euclidean geometry for the unaltered original compositions. The Aitchison geometry faces an immediate problem in this data application because a substantial proportion of mood compositions is zero, leaving the log-ratio undefined; this poses no problems for the proposed square-root transformation approach. Though the compositional simplex can be identically embedded into the Euclidean space and endowed with the Euclidean geometry, this approach yields a geodesic distance that equally emphasizes the differences in the entries with large or small magnitude. In many applications, it is more sensible to attach higher importance to the small entries, and this is effectively

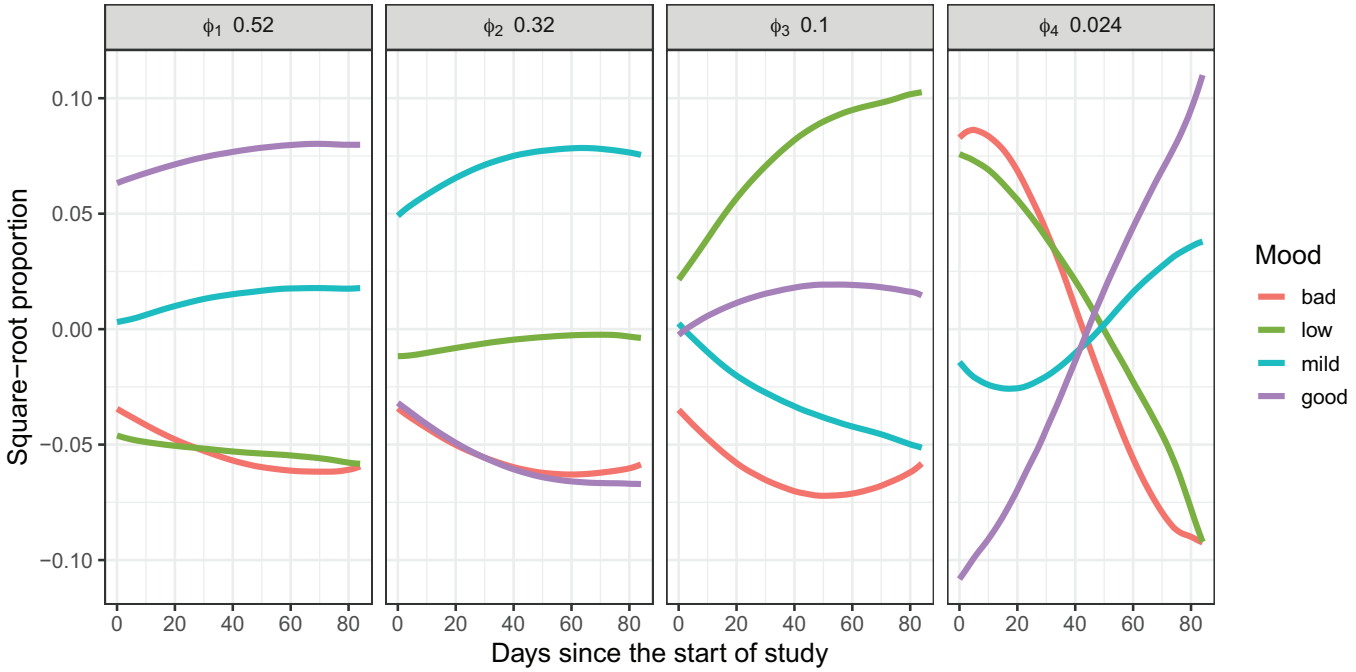

FIGURE 4 The first four eigenfunctions for the longitudinal mood composition data in the square root transformation scale, with fraction of variation explained (FVE) displayed in the panel subtitles. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

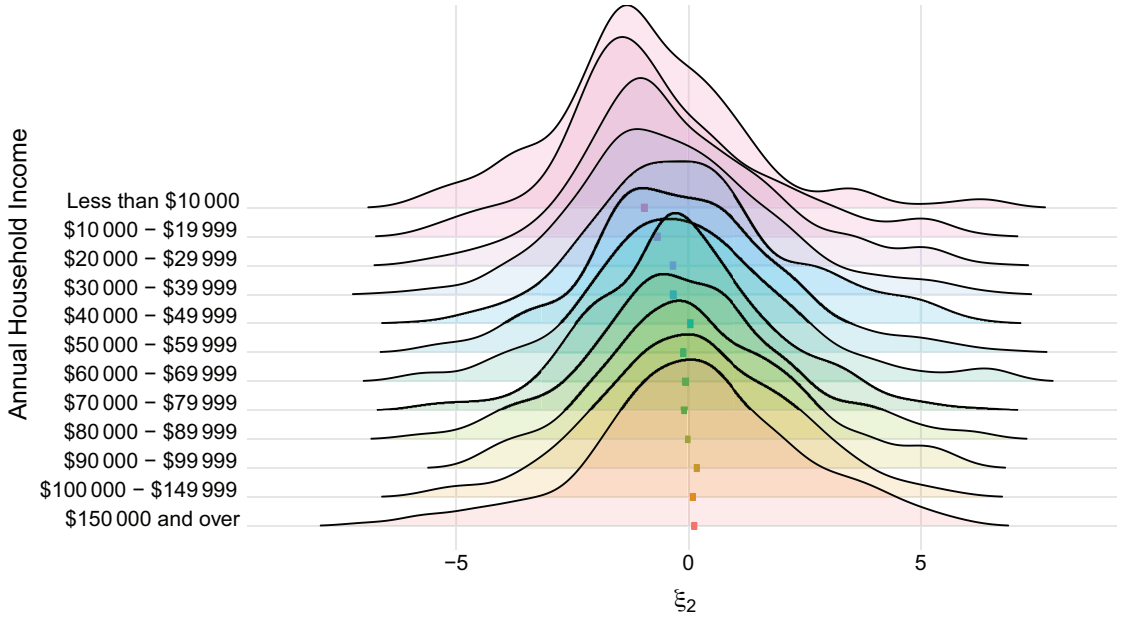

FIGURE 5 The distributions of the second Riemannian principal component score, encoding emotion stability, visualized as densities in dependence on the annual household income in 2008. Colored dots indicate the mean of this score for each income group. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

achieved by the square-root transformation and also by the log transformation.

Bandwidths were selected by GCV as $h_\mu = 17.9$ and $h_\Gamma = 35.8$ days with the Epanechnikov kernel. The fitted mood composition trajectories are displayed in the left panel of Figure 3 for four randomly selected subjects, where the solid dots denote the reported moods and are slightly jittered vertically if they overlap, and

dashed curves denote the fitted trajectories. The fits are obtained with $K = 8$ components, selected according to the FVE criterion (8) with threshold $\gamma = .99$, which is a reasonable choice in view of the large sample size. As the self-reported moods contain substantial aberrations from smooth trajectories that we view as noise, the fitted trajectories do not go through the raw observations. The mean trajectory is displayed in the right panel of Figure 3, indicating that the emotional well-being of subjects tends to deteriorate as the period of unemployment lengthens, with an overall increase in the proportion of bad mood and a decrease in the proportion of good mood.

The first four eigenfunctions for mood composition trajectories are shown in Figure 4, where the first eigenfunction corresponds to the overall contrast between neutral-to-positive mood (good and mild) and negative moods (low and bad); the second eigenfunction represents emotional stability, which is a contrast between more neutral moods and extreme emotions (good and bad); the third eigenfunction corresponds to a shift of mood compositions to more positive moods, namely, from bad to low and from mild to good; the fourth eigenfunction encodes an increase of positive feelings and a decrease of negative ones over time. Here it is important to note that the sign of the eigenfunctions is arbitrary and could be reversed. The first four eigenfunctions together explain 95% of the total variation.

To demonstrate that the scores obtained from the proposed approach are useful for downstream tasks such as regression, we explored the association between the second Riemannian principal component score ξ_{i2}, corresponding to the proportion of extreme moods, and annual household income in 2008, a measure of financial stability. Collecting these scores for all subjects, we constructed kernel density estimates for the ξ_{i2} within each income category; see Figure 5. Participants with higher household income before losing their job and thus higher financial stability tend to have higher emotion stability, as demonstrated by the right-shifted distributions of ξ_{i2} and larger means (colored dots). The relationship between prior income and emotional stability appears to be nonlinear, especially for lower income groups.

5 | CONCLUDING REMARKS

While the proposed RPACE approach has been found to perform very well in the data examples and simulation, and will provide a valuable tool for longitudinal studies with complex data, its utility finds limits for extremely sparse longitudinal designs with an average of around two measurements per subject. Such extremely sparse designs do occur in practical applications (Dai *et al.*, 2019). In such cases, mean and linearized covariance functions can still be reasonably estimated if the number of subjects is large, but the recovery of individual trajectories is unstable. Further limitations are encountered for manifolds with high curvature where local linear approximations work less well and for stratified or infinite-dimensional manifolds. To address and overcome these limits will be left for future research.

ACKNOWLEDGMENTS

We thank the reviewers for their constructive comments. This research was supported by NSF grants DMS-1712864 and DMS-2014626. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are publicly available at adni.loni.usc.edu and opr.princeton.edu/archive/njui/.

ORCID

Xiongtao Dai https://orcid.org/0000-0002-6996-5930

REFERENCES

Afsari, B. (2011) Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139, 655–673.

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Anirudh, R., Turaga, P., Su, J. and Srivastava, A. (2017) Elastic functional coding of Riemannian trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 922–936.

Badhwar, A., Tam, A., Dansereau, C., Orban, P., Hoffstaedter, F. and Bellec, P. (2017) Resting-state network dysfunction in Alzheimer's disease: a systematic review and meta-analysis. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 8, 73–85.

Bhattacharya, A. and Bhattacharya, R. (2012) *Nonparametric Inference on Manifolds: With Applications to Shape Spaces*, volume 2. Cambridge: Cambridge University Press.

Bhattacharya, R. and Patrangenaru, V. (2003) Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *The Annals of Statistics*, 31, 1–29.

Bosq, D. (2000) *Linear Processes in Function Spaces*. New York: Springer.

Buckner, R.L., Sepulcre, J., Talukdar, T., Krienen, F.M., Liu, H., Hedden, T. et al. (2009) Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *The Journal of Neuroscience*, 29, 1860–1873.

Chen, D. and Müller, H. (2012) Nonlinear manifold representations for functional data. *Annals of Statistics*, 40, 1–29.

Chen, K. and Lei, J. (2015) Localized functional principal component analysis. *Journal of the American Statistical Association*, 110, 1266–1275.

Cornea, E., Zhu, H., Kim, P. and Ibrahim, J.G. (2017) Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 463–482.

Dai, X., Hadjipantelis, P., Wang, J.-L., Deoni, S.C. and Müller, H.-G. (2019) Longitudinal associations between white matter maturation and cognitive development across early childhood. *Human Brain Mapping*, 40, 4130–4145.

Dai, X. and Müller, H.-G. (2018) Principal component analysis for functional data on Riemannian manifolds and spheres. *Annals of Statistics*, 46, 3334–3361.

Deoni, S.C., O'Muircheartaigh, J., Elison, J.T., Walker, L., Doernberg, E., Waskiewicz, N., Dirks, H., Piryatinsky, I., Dean, D.C. and Jumbe, N. (2016) White matter maturation profiles through early childhood predict general cognitive ability. *Brain Structure and Function*, 221, 1189–1203.

Dubey, P. and Müller, H.-G. (2019) Fréchet analysis of variance for random objects. *Biometrika*, 106, 803–821.

Dubey, P. and Müller, H.-G. (2020) Functional models for time-varying random objects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 275–327.

Dubin, J.A. and Müller, H.-G. (2005) Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100, 872–881.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Fang, K., Kotz, S. and Ng, K. (1990) *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall.

Fréchet, M. (1948) Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, 10, 215–310.

Ginestet, C.E., Li, J., Balachandran, P., Rosenberg, S. and Kolaczyk, E.D. (2017) Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11, 725–750.

Grenander, U. (1950) Stochastic processes and statistical inference. *Arkiv för Matematik*, 1, 195–277.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006) Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34, 1493–1517.

Hart, B., Cribben, I., Fiecas, M. and Alzheimer's Disease Neuroimaging Initiative, (2018) A longitudinal model for functional connectivity networks using resting-state fMRI. *NeuroImage*, 178, 687–701.

Horvath, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications*. New York: Springer.

Hsing, T. and Eubank, R. (2015) *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Hoboken: Wiley.

Kleffe, J. (1973) Principal components of random variables with values in a separable Hilbert space. *Statistics*, 4, 391–406.

Kong, D., Xue, K., Yao, F. and Zhang, H.H. (2016) Partially functional linear regression in high dimensions. *Biometrika*, 103, 147–159.

Krueger, A.B. and Mueller, A. (2011) Job search, emotional well-being, and job finding in a period of mass unemployment: evidence from high-frequency longitudinal data. *Brookings Papers on Economic Activity*, 2011, 1–57.

Kundu, S., Lukemire, J., Wang, Y. and Guo, Y. (2019) A novel joint brain network analysis using longitudinal Alzheimer's disease data. *Scientific Reports*, 9, 1–18.

Laumann, T.O., Snyder, A.Z., Mitra, A., Gordon, E.M., Gratton, C., Adeyemo, B., Gilmore, A.W., Nelson, S.M., Berg, J.J., Greene, D.J., McCarthy, J.E., Tagliazucchi, E., Laufs, H., Schlaggar, B.L., Dosenbach, N.U.F., and Petersen, S.E. (2017) On the stability of BOLD fMRI correlations. *Cerebral Cortex*, 27, 4719–4732.

Lee, J.M. (1997) *Riemannian Manifolds: An Introduction to Curvature*. New York: Springer-Verlag.

Li, Y. and Hsing, T. (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, 38, 3321–3351.

Lin, L., St.Thomas, B., Zhu, H. and Dunson, D.B. (2017) Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association*, 112, 1261–1273.

Lin, Z. and Yao, F. (2019) Intrinsic Riemannian functional data analysis. *Annals of Statistics*, 47, 3533–3577.

Nash, J. (1956) The imbedding problem for Riemannian manifolds. *Annals of Mathematics*, 63, 20–63.

Patrangenaru, V., Bubenik, P., Paige, R.L. and Osborne, D. (2018) Topological data analysis for object data. arXiv preprint, arXiv:1804.10255.

Pennec, X., Fillard, P. and Ayache, N. (2006) A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66, 41–66.

Petersen, A. and Müller, H.-G. (2019) Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, 47, 91–719.

Scealy, J. and Welsh, A. (2011) Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 351–375.

Telschow, F.J.E., Pierrynowski, M.R. and Huckemann, S.F. (2019) Confidence tubes for curves on SO(3) and identification of subject-specific gait change after kneeling. arXiv preprint, arXiv: 1909.06583.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and its Application*, 3, 257–295.

Worsley, K.J., Chen, J.-I., Lerch, J. and Evans, A.C. (2005) Comparing functional connectivity via thresholding correlations and singular value decomposition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 913–920.

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.

Yuan, Y., Zhu, H., Lin, W. and Marron, J.S. (2012) Local polynomial regression for symmetric positive definite matrices. *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, 74, 697–719.

Zhang, X. and Wang, J.L. (2016) From sparse to dense functional data and beyond. *The Annals of Statistics*, 44, 2281–2321.

Zhang, X. and Wang, J.-L. (2018) Optimal weighting schemes for longitudinal and functional data. *Statistics & Probability Letters*, 138, 165–170.

SUPPORTING INFORMATION

Web Appendices for the simulation referenced in Section 1, proofs and additional discussions for the theoretical results referenced in Section 3, and links to the data and code for reproducing the data applications are available with this paper at the Biometrics website on Wiley Online Library. An R package implementation is available at https://github.com/CrossD/RFPCA.

How to cite this article: Dai X, Lin Z, Müller Hans-Georg. Modeling sparse longitudinal data on Riemannian manifolds. *Biometrics*. 2020;1–14. https://doi.org/10.1111/biom.13385