

# NONLINEAR MANIFOLD REPRESENTATIONS FOR FUNCTIONAL DATA

BY DONG CHEN AND HANS-GEORG MÜLLER<sup>1</sup>

*University of California, Davis*

For functional data lying on an unknown nonlinear low-dimensional space, we study manifold learning and introduce the notions of manifold mean, manifold modes of functional variation and of functional manifold components. These constitute nonlinear representations of functional data that complement classical linear representations such as eigenfunctions and functional principal components. Our manifold learning procedures borrow ideas from existing nonlinear dimension reduction methods, which we modify to address functional data settings. In simulations and applications, we study examples of functional data which lie on a manifold and validate the superior behavior of manifold mean and functional manifold components over traditional cross-sectional mean and functional principal components. We also include consistency proofs for our estimators under certain assumptions.

**1. Introduction.** Nonlinear dimension reduction methods, such as locally linear embedding [28], isometric mapping [31] and Laplacian eigenmaps [2], have been successfully applied to image data in recent years. A commonly used example is the analysis of photos of a sculpture face taken under different angles and lighting conditions. The number of pixels of these images is huge, but their structure only depends on a few variables related to angle and lighting conditions. It is then advantageous to treat the observed image data as a manifold that is approximately *isomorphic* to a low-dimensional Euclidean space.

Unlike shape analysis [21] and the recent diffusion tensor imaging [17], where it is assumed that the form of the manifold is known a priori, nonlinear dimension reduction methods usually are manifold-learning procedures, where the manifold is not known but it is assumed that it possesses certain

---

Received June 2011; revised September 2011.

<sup>1</sup>Supported in part by NSF Grants DMS-08-06199 and DMS-11-04426.

*AMS 2000 subject classifications.* 62H25, 62M09.

*Key words and phrases.* Functional data analysis, modes of functional variation, functional manifold components, dimension reduction, smoothing.

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *The Annals of Statistics*, 2012, Vol. 40, No. 1, 1–29. This reprint differs from the original in pagination and typographic detail.

features which are preserved in the observed data. For instance, locally linear embedding preserves the manifold local linear structure while isometric mapping preserves geodesic distance. Their inherent flexibility predisposes these methods for extensions to functional data, where one rarely would have prior information available about the nature of the underlying manifold.

Our goal is to explore manifold representations of functional data. Which observed sets of functions are likely to lie on a low-dimensional manifold? And how should this be taken into consideration? In contrast to multivariate data, functional data are recorded on a time or location domain, and commonly are assumed to consist of sets of smooth random functions. Auspicious examples where functional manifold approaches may lead to improved representations include time-warped functional data [12, 33], density functions [23], and functional data with pre-determined and interpretable modes [18]. In such situations, the established linear functional approaches, such as cross-sectional mean and functional principal component analysis (FPCA) often fail to represent the functional data in a parsimonious, efficient and interpretable way. Manifold approaches are expected to be especially useful to represent functional data inherently lying on a low-dimensional nonlinear space.

In this paper, we develop a framework for modeling  $L^2$  functions on unknown manifolds and propose pertinent notions, such as manifold mean, manifold modes of functional variation and functional manifold components, as elements of a functional manifold component analysis (FMCA). Manifold means complement notions of a specifically modified functional mean, such as the “structural mean” [22]. A major motivation for this proposal is that functional principal component plots, for example, displaying second versus first component, are quite often found to exhibit “horseshoe” shapes, that is, nonlinear dependence in the presence of uncorrelatedness (as principal components by definition are always uncorrelated). An example of this “horseshoe shape” is provided by the Berkeley growth data (see upper right panel of Figure 5). In such situations, one may wish to “unwrap” the “horseshoe” into linear structures by techniques similar to those used in nonlinear dimension reduction. When attempting to “unwrap” functional data, one encounters specific difficulties: Often the underlying smooth functions are not directly observed, but instead need to be inferred from a limited number of noise-contaminated measurements that contain the available information for each subject in the sample. To address these problems, we develop a modified ISOMAP [31] procedure, by adding a data-adaptive penalty to the empirical geodesic distances, and employ local smoothing to recover the manifold.

The paper is organized in the following way. In Section 2, we describe what we mean by a functional manifold, manifold mean, manifold modes of functional variation and functional manifold components. We develop corresponding estimates in Section 3 and discuss their asymptotic properties

in Section 4. Sections 5 and 6 are devoted to illustrations of the proposed methodology for both simulated and real data. Detailed proofs can be found in an online supplement [5].

## 2. Manifolds in function space.

*2.1. Preliminaries.* A manifold  $\mathcal{M}$  can be expressed in terms of an atlas consisting of a group of charts  $(U_\alpha, \varphi_\alpha)$ , where  $U_\alpha$  are open sets covering  $\mathcal{M}$  and  $\varphi_\alpha$ , the coordinate maps, map the corresponding  $U_\alpha$  onto an open subset of  $\mathbb{R}^d$ . Additional assumptions on  $\varphi_\alpha$  are usually imposed in order to study the structure of  $\mathcal{M}$  [8, 16].

In this paper, we only consider “simple” functional manifolds  $\mathcal{M}$  in  $L^2$  space, where  $\mathcal{M}$  is *isomorphic* to a subspace of the Euclidean space, that is, the manifold can be represented by a coordinate map  $\varphi: \mathbb{R}^d \rightarrow \mathcal{M} \subset L^2$ , such that  $\varphi$  is bijective, and both  $\varphi, \varphi^{-1}$  are continuous, in the sense that if  $\boldsymbol{\theta}_n, \boldsymbol{\theta} \in \mathbb{R}^d$  and  $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\| \rightarrow 0$ ,  $\|\varphi(\boldsymbol{\theta}_n) - \varphi(\boldsymbol{\theta})\|_{L^2} \rightarrow 0$ ; if  $x_n, x \in \mathcal{M}$  and  $\|x_n - x\|_{L^2} \rightarrow 0$ ,  $\|\varphi^{-1}(x_n) - \varphi^{-1}(x)\| \rightarrow 0$ . Here,  $d$  is the intrinsic dimension of the manifold  $\mathcal{M}$ . Such “simple” manifold settings have been commonly considered in the dimension reduction literature, for example in [31].

For a continuous curve defined on the manifold  $\gamma: [0, 1] \rightarrow \mathcal{M}$ , define the length operator

$$(2.1) \quad L(\gamma) = \sup \sum_{i=0}^{n-1} \|\gamma(s_{i+1}) - \gamma(s_i)\|_{L^2},$$

where the supremum is taken over all partitions of the interval  $[0, 1]$  with arbitrary break points  $0 = s_0 < s_1 < \dots < s_n = 1$ . We call  $\varphi$  an isometric map if  $L(\gamma) = L(\varphi^{-1} \circ \gamma)$  for any continuous  $\gamma$ , where  $L(\varphi^{-1} \circ \gamma)$  is similarly defined as in (2.1) with the  $L^2$  norm replaced by the Euclidean norm. We say  $\mathcal{M}$  is an isometric manifold if there exists an isometric coordinate map  $\varphi$ . The isometry assumption is pragmatically desirable and can be found in many approaches [9, 31]. Conditions under which isometry holds for image data are discussed in [10].

We use the notation  $\psi \equiv \varphi^{-1}$  and refer to  $\psi$  as the representation map. The manifold  $\mathcal{M}$  is naturally equipped with the  $L^2$  distance, which, due to the nonlinearity of  $\mathcal{M}$ , is not an adequate metric [31]. More useful is the geodesic distance

$$(2.2) \quad d_g(x_1, x_2) = \inf \{L(\gamma) : \gamma(0) = x_1, \gamma(1) = x_2\},$$

where the infimum is taken over all continuous paths  $\gamma$  on  $\mathcal{M}$ . The geodesic distance is the length of the shortest path on  $\mathcal{M}$  connecting the two points, and therefore is adapted to  $\mathcal{M}$ .

*2.2. Manifold mean and manifold modes of variation.* Suppose  $\mathcal{M}$  is a functional manifold of intrinsic dimension  $d$  and  $\psi$  is a representation

map for  $\mathcal{M}$ . Define, with respect to a probability measure  $Q$  in  $\mathbb{R}^d$ ,

$$(2.3) \quad \boldsymbol{\mu} = \mathbb{E}\{\psi(X)\}, \quad \mu^{\mathcal{M}} = \psi^{-1}(\boldsymbol{\mu}),$$

where  $\boldsymbol{\mu}$  is the mean in the  $d$ -dimensional representation space, and  $\mu^{\mathcal{M}}$  is the manifold mean in  $L^2$  space. If  $\mathcal{M}$  is isometric, the manifold mean  $\mu^{\mathcal{M}}$  is uniquely defined for all isometric representation maps, as the following results shows.

**PROPOSITION 1.** *Suppose the random function  $X$  lies on a functional manifold  $\mathcal{M}$  of intrinsic dimension  $d$  and  $\psi$  is a representation map for  $\mathcal{M}$ . If  $\psi$  is isometric, the manifold mean  $\mu^{\mathcal{M}}$  in (2.3) has the following alternative expression:*

$$(2.4) \quad \mu^{\mathcal{M}} = \arg \min_{x \in \mathcal{M}} \mathbb{E} d_g^2(x, X),$$

where  $d_g$  denotes the geodesic distance defined in (2.2).

The expected value in equation (2.4) is with respect to the probability measure that is induced by the map  $\varphi$ ; see also [3]. Equation (2.4) defines the Fréchet mean for geodesic distance  $d_g(\cdot, \cdot)$ , and therefore does not depend on the choice of the isometric map  $\psi$ . The motivation to consider the manifold mean is that the traditional cross-sectional mean for functional data in  $L^2$  has significant drawbacks as a measure of location when the data indeed lie on a nonlinear functional manifold. Estimates of  $L^2$  means, obtained by averaging observed sample curves, can be far away from the data cloud in such situations, and therefore do not represent the data in a meaningful way. Going beyond the mean, one encounters analogous problems when linearly representing such random functions in an  $L^2$  basis, such as the Fourier, B spline or eigenfunction basis.

Consider random functions  $X \in L^2(\mathcal{T})$  defined on a bounded domain  $\mathcal{T}$ . With  $\mu(t) = \mathbb{E}X(t)$  and  $G(t, s) = \text{Cov}(X(t), X(s))$ , according to Mercer's theorem [1], if the covariance function  $G(t, s)$  is jointly continuous in  $t, s$ , there is an orthonormal expansion of  $G(t, s)$  in terms of the eigenvalues  $\{\lambda_k : k \geq 1\}$  (ordered nonincreasingly) and associated eigenfunctions  $\{\phi_k : k \geq 1\}$ ,

$$(2.5) \quad G(t, s) = \sum_{k=1}^{\infty} \lambda_k \phi_k(t) \phi_k(s), \quad t, s \in \mathcal{T}.$$

By the Hilbert–Schmidt theorem [14, 27],  $X$  can be expressed in terms of the so-called Karhunen–Loève representation,

$$(2.6) \quad X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t), \quad t \in \mathcal{T},$$

$$\xi_k = \int_{\mathcal{T}} (X(t) - \mu(t)) \phi_k(t) dt,$$

where the  $\xi_k$  are uncorrelated random variables with mean 0 and variance  $\lambda_k$ , known as functional principal components (FPCs).

In the manifold case, the FPCs intrinsically lie on a  $d$ -dimensional manifold. Therefore, we expect that the FPCs do not provide a parsimonious representation of  $X$ . A better adapted and more compact representation can be obtained through nonlinear manifold modes of functional variation that are defined below. The established eigenfunction-based modes of functional variation [4, 19] are

$$(2.7) \quad X_{j,\alpha} = \mu + \alpha \lambda_j^{1/2} \phi_j, \quad j = 1, 2, \dots, \alpha \in \mathbb{R},$$

where factors  $\lambda_j^{1/2}$  standardize the scale for different  $j$  and the functional variation in the direction of eigenfunction  $\phi_j$  is visualized by the changing of functional shapes as  $\alpha$  varies. However, when the functional data lie on a manifold, neither  $\mu$  nor  $X_{j,\alpha}$  may belong to  $\mathcal{M}$ , so that these linear modes will not provide a sensible description of the variation in the data.

To address this problem, we define functional manifold component (FMC) vectors  $\mathbf{e}_j \in \mathbb{R}^d$ ,  $j = 1, \dots, d$ , by the eigenvectors of the covariance matrix of  $\psi(X) \in \mathbb{R}^d$ , that is,

$$(2.8) \quad \text{Cov}(\psi(X)) = \sum_{j=1}^d \lambda_j^{\mathcal{M}} \mathbf{e}_j \mathbf{e}_j^T,$$

where  $\lambda_1^{\mathcal{M}} \geq \dots \geq \lambda_d^{\mathcal{M}}$  are the eigenvalues of  $\text{Cov}(\psi(X))$ . The manifold modes of functional variation are

$$(2.9) \quad X_{j,\alpha}^{\mathcal{M}} = \psi^{-1}(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{1/2} \mathbf{e}_j), \quad j = 1, \dots, d, \alpha \in \mathbb{R},$$

where  $\boldsymbol{\mu}$  is the mean in the  $d$ -dimensional representation space according to measure  $Q$ , as given in (2.3). A distinct advantage of manifold-based modes of functional variation over the principal component based version (2.7) is that in (2.9) only finitely many modes are needed, while (2.7) requires potentially infinitely many components. The manifold modes  $X_{j,\alpha}^{\mathcal{M}}$  are unique for the case of isometric  $\mathcal{M}$ , as shown in the following.

**PROPOSITION 2.** *Suppose  $\psi$  and  $\tilde{\psi}$  are two isometric representation maps for a functional manifold  $\mathcal{M}$  of intrinsic dimension  $d$ . Let  $X_{j,\alpha}^{\mathcal{M}}$  be the  $j$ th manifold mode defined in (2.9) based on representation map  $\psi$ , and  $\tilde{X}_{j,\alpha}^{\mathcal{M}}$  be the  $j$ th manifold mode using map  $\tilde{\psi}$ . Then  $X_{j,\alpha}^{\mathcal{M}} = \tilde{X}_{j,\alpha}^{\mathcal{M}}$  for all  $\alpha \in \mathbb{R}$  and  $1 \leq j \leq d$ , if the eigenvalues of  $\text{Cov}(\psi(X))$  and of  $\text{Cov}(\tilde{\psi}(X))$  are of multiplicity one.*

For each  $X \in \mathcal{M}$ , given the representation map  $\psi$ ,  $X$  can be uniquely represented (due to the bijectivity of  $\psi$ ) by a vector  $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_d) \in \mathbb{R}^d$

$$(2.10) \quad X = \psi^{-1}\left(\boldsymbol{\mu} + \sum_{j=1}^d \vartheta_j \mathbf{e}_j\right), \quad \vartheta_j = \langle \psi(X) - \boldsymbol{\mu}, \mathbf{e}_j \rangle, \quad j = 1, \dots, d,$$

where  $\boldsymbol{\mu}$  and  $\mathbf{e}_j$  are defined in (2.3) and (2.8), respectively,  $\langle \cdot, \cdot \rangle$  is the inner product in  $\mathbb{R}^d$  and  $\vartheta_j$  are uncorrelated r.v.s with mean 0 and variance  $\lambda_j^{\mathcal{M}}$ . We call  $\vartheta_j$  the functional manifold components (FMCs) in the representation space.

**3. Estimating functional manifolds.** Suppose we observe  $\{Y_{ij} : 1 \leq i \leq n; 1 \leq j \leq n_i\}$  which are noise-contaminated measurements made on  $n$  independent realizations  $X_i$  of a random function  $X \in \mathcal{M}$ , according to the data model

$$Y_{ij} = X_i(t_{ij}) + \varepsilon_{ij}.$$

Here the  $t_{ij}$  are the time points where the functions are sampled, and the  $\varepsilon_{ij} \in \mathbb{R}$  are i.i.d. errors with mean 0 and variance  $\sigma^2$ . A first task is to find an approximation  $\hat{\psi}$  to the representation map  $\psi$  based on the observed  $Y_{ij}$ . We also require the inverse  $\hat{\psi}^{-1}$ . Prior knowledge about the data may suggest a specific form for  $\psi$  [18], or one may have direct observations of  $\psi(X_i)$ . But in general, the representation map  $\psi$  is unknown and needs to be determined from the data.

**3.1. Inferring  $d$ -dimensional manifold representations.** Following [31], we use the pairwise distances between observed data to obtain a map  $\psi$  that preserves the geodesic distances. Alternative approaches include LLE [28] and Laplacian eigenmaps [2]. While these methods have been developed for multivariate data, they can be adapted to functional data in a two-step procedure as follows.

In a first step, given an intrinsic dimension  $d$  of  $\mathcal{M}$ , adopt the proposal of [31] to obtain the function  $\psi : L^2 \rightarrow \mathbb{R}^d$  only at the sample points  $\{X_1, \dots, X_n\}$ , where  $X_i \in L^2$ , by

$$(3.1) \quad \hat{\psi} = \arg \min_{(\psi(X_1), \dots, \psi(X_n))} \sum_{i,j=1}^n \{\|\psi(X_i) - \psi(X_j)\| - d_g(X_i, X_j)\}^2.$$

Here,  $d_g(\cdot, \cdot)$  is the geodesic distance (2.2) and the minimum is taken over the vectors  $\psi(X_i) \in \mathbb{R}^d, i = 1, \dots, n$ , formed by the values of  $\psi$  on the functions  $X_i$ , that is, the goal is to find  $n$  vectors  $\hat{\psi}(X_i) \in \mathbb{R}^d, i = 1, \dots, n$ , that minimize (3.1). For this, one needs to estimate the geodesic distances, and then the minimizer  $\hat{\psi}(X_i)$  is obtained by multidimensional scaling (MDS) based on estimates of  $d_g(X_i, X_j)$  [6]. Our asymptotic results pertain to a second step, where the assumed smoothness of  $\psi$  is invoked to obtain global estimates for  $\hat{\psi}$ , as described in Section 3.2. As for  $\hat{\psi}(X_i), i = 1, \dots, n$ , as determined by (3.1), we assume that the minimization in (3.1) provides values on or defines the target manifold at the sample points, that is, that  $\hat{\psi}(X_i) = \psi(X_i), i = 1, \dots, n$ , or alternatively, that  $v_n = \hat{\psi}(X_i) - \psi(X_i) \rightarrow 0$ .

In order to approximate geodesic distances  $d_g(X_i, X_j)$ , we first aim at estimates of the  $L^2$  distances  $\|X_i - X_j\|_{L^2}$ . For this purpose, the Karhunen–Loève representation (2.6) can be used to obtain fitted curves,

$$(3.2) \quad \hat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k \leq K} \hat{\xi}_{ik} \hat{\phi}_k(t).$$

Here,  $\hat{\mu}(t)$  and  $\hat{G}(t, s)$  are first obtained by applying local linear one-dimensional and two-dimensional smoothers to the pooled data; then eigenfunctions  $\hat{\phi}_k(t)$  and eigenvalues  $\hat{\lambda}_k$  are extracted by classical vector spectral analysis applied to a discretized version of the estimate  $\hat{G}(t, s)$  of the covariance surface  $G(t, s) = \text{Cov}(X(t), X(s))$ ; and then the FPCs  $\xi_{ik}$  are approximated by discretizing integrals

$$(3.3) \quad \hat{\xi}_{ik} = \sum_{j=2}^{n_i} \{Y_{ij} - \hat{\mu}(t_{ij})\} \hat{\phi}_k(t_{ij}) (t_{ij} - t_{i,j-1})$$

or alternatively by conditional expectation (for details on these steps, see [34]),

$$(3.4) \quad \hat{\xi}_{ik} = \hat{\lambda}_k \hat{\phi}_{ik}^T \hat{\Sigma}_{\mathbf{Y}_i}^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i),$$

where  $\hat{\phi}_{ik} = (\hat{\phi}_k(t_{i1}), \dots, \hat{\phi}_k(t_{in_i}))$ ,  $(\hat{\Sigma}_{\mathbf{Y}_i})_{jl} = \hat{G}(t_{ij}, t_{il}) + \hat{\sigma}^2 \mathbf{I}$ ,  $1 \leq j, l \leq n_i$ ,  $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}(t_{i1}), \dots, \hat{\mu}(t_{in_i}))$ , and  $\sigma^2$  is estimated from the difference between empirical variances of  $Y_{ij}$  and  $\hat{G}(t, s)$ . The conditioning method (3.4) is the only available option if the data are sparsely sampled. To ensure that a sufficiently large number of components is included in the truncated expansion (3.2), one may choose  $K$  by requiring a large fraction of variance explained (FVE), that is,

$$(3.5) \quad K = \min_k \left\{ k : \frac{\sum_{l \leq k} \hat{\lambda}_l}{\sum_{l=1}^{\infty} \hat{\lambda}_l} \geq 1 - \alpha \right\}$$

for, say,  $\alpha = 0.05$ , where the  $\hat{\lambda}_l$  are estimates of the eigenvalues  $\lambda_l$  in (2.5). The resulting  $L^2$  distances are  $\|\hat{X}_i^K - \hat{X}_j^K\|_{L^2} = \{\sum_{k=1}^K (\hat{\xi}_{ik} - \hat{\xi}_{jk})^2\}^{1/2}$ .

Note that alternatively to representation (3.2), one can also directly apply local constant or local linear smoothing to obtain smooth trajectories in the case of dense and balanced designs, for example, using Nadaraya–Watson kernel estimators,

$$(3.6) \quad \tilde{X}_i(t) = \frac{\sum_{j=1}^{n_i} \kappa_1(h_1^{-1}(t_{ij} - t)) Y_{ij}}{\sum_{j=1}^{n_i} \kappa_1(h_1^{-1}(t_{ij} - t))},$$

where  $\kappa_1$  and  $h_1$  are smoothing kernel and bandwidth. For the smoothing kernel one can use any standard kernel such as the standard Gaussian density

function or the Epanechnikov kernel, while in practice  $h_1$  may be chosen by cross-validation or generalized cross-validation.

Then the pairwise  $L^2$  distances are simply  $\|\tilde{X}_i - \tilde{X}_j\|_{L^2}$ . We will not explicitly explore this alternative smoothing approach in our theoretical analysis, but note that essentially the same results as those reported below hold for this alternative approach, by minor extensions of our arguments. In the implementations (simulation and data analysis), we use both approaches (3.2) and (3.6). The estimated random trajectories, obtained through (3.2) or (3.6), generally are not lying on the manifold  $\mathcal{M}$ , as they are merely approximations to the true unknown functions, due to additional noise and discrete sampling of the random trajectories. However, these estimates, owing to their consistency, will fall inside a small  $L^2$ -neighborhood around  $\mathcal{M}$ . Asymptotic properties are discussed in Section 4.

Since the geodesic is the shortest path connecting points on a manifold, one may first connect the points inside small  $L^2$  neighborhoods and then define the path between two far away points by moving along these small neighborhoods, and then find the geodesic by the shortest path connecting through such neighborhoods. This is essentially the idea of the ISOMAP algorithm [31]. The performance of this method however proved somewhat unstable in our applications, as functional data typically must be inferred from discretized and noisy observations of underlying smooth trajectories and therefore do not exactly lie on the manifold, as is assumed in ISOMAP.

In such situations, due to random scattering of the data around the manifold, the shortest path found by the ISOMAP criterion may pass through “empty areas” outside the proper data cloud. This problem can be effectively addressed by modifying the ISOMAP criterion, by additionally penalizing against paths that include sections situated within “empty regions” with few neighboring data points. Density-penalized geodesics are characterized by sequences of  $L^2$  functions  $(W_1, W_2, \dots, W_m)$  from the starting point  $W_s = W_1$  to the end point  $W_e = W_m$  of the geodesic, where each of the  $W_j$  stands for one of the observed functions  $X_i$  (with unrelated index), and are defined as

$$(3.7) \quad S(W_s, W_e) = \arg \min_{W_2, \dots, W_{m-1}} \left\{ \sum_{i=1}^{m-1} \|W_i - W_{i+1}\|_{L^2} (1 + P_\delta(W_i, W_{i+1})) : \right. \\ \left. \|W_i - W_{i+1}\|_{L^2} < \varepsilon \right\}.$$

Here the parameter  $\varepsilon$  limits the step length, and the penalty function  $P_\delta$  is determined by the density of the data cloud around  $W_i$  and  $W_{i+1}$ ,

$$P_\delta(W_i, W_{i+1}) = \rho_{i,i+1}^{-2} I(\rho_{i,i+1} < \delta),$$



where  $\rho_{i,i+1} = \min\{\#\{W_j : \|W_j - W_i\|_{L^2} < \varepsilon\}, \#\{W_j : \|W_j - W_{i+1}\|_{L^2} < \varepsilon\}\}$  and  $\#$  denotes the cardinality of a set. By selecting the parameter  $\delta$ , one can control the threshold of the local density of points, below which the penalty  $P_\delta$  kicks in. The ISOMAP algorithm corresponds to the special case where  $\delta = 0, P_\delta = 0$ .

The choice  $\delta > 0$  leads to “penalized ISOMAP” or P-ISOMAP, where the penalty parameter  $\delta$  may be selected data-adaptively by cross-validation. The choice of  $\delta$  and also of the step size parameter  $\varepsilon$  is discussed in Section 3.3. If the manifold is very smooth, a large  $\varepsilon$  and small  $m$  will lead to a sufficiently good estimate of the geodesic distance. A detailed discussion of the convergence of the estimated geodesics in the framework of ISOMAP can be found at <http://isomap.stanford.edu/BdSLT.pdf>. For the proposed P-ISOMAP, we implement the minimization of  $S(W_s, W_e)$  by Dijkstra’s algorithm, which selects  $m$  and the geodesic paths ( $W_s = W_1, W_2, \dots, W_{m-1}, W_e = W_m$ ). The resulting estimated geodesic distance is

$$(3.8) \quad \hat{d}_g(W_s, W_e) = \sum_{j=1}^{m-1} \|\hat{W}_j - \hat{W}_{j+1}\|_{L^2},$$

where  $\hat{W}_j = \tilde{W}_j$  or  $\hat{W}_j^K$ , depending on which preliminary approximation is used for  $W_j$ . Once these distances have been determined, an application of MDS yields  $\hat{\psi}(X_i)$ , in the same way as in the standard ISOMAP method.

*3.2. Obtaining the global map and representing sample trajectories.* For any location  $\theta \in \mathbb{R}^d$ , we find  $\hat{\psi}^{-1}(\theta)$  by local weighted averaging, that is,

$$(3.9) \quad \hat{\psi}^{-1}(\theta) = \frac{\sum_i \kappa(H^{-1}(\hat{\psi}(X_i) - \theta)) \hat{X}_i}{\sum_i \kappa(H^{-1}(\hat{\psi}(X_i) - \theta))},$$

where  $\kappa$  is a  $d$ -dimensional kernel, like the Epanechnikov kernel  $\kappa(u_1, \dots, u_d) = (\frac{3}{4})^d \prod_{k=1}^d \{(1 - u_k^2) \mathbf{I}(|u_k| < 1)\}$ , with  $H = h \mathbf{I}_{d \times d}$  for a suitably chosen bandwidth  $h$ ,  $\hat{X}_i$  could be either  $\tilde{X}_i$  as in (3.6) or  $\hat{X}_i^K$  as in (3.2), and  $\hat{\psi}(X_i)$  is defined after (3.8). We use cross-validation to select  $h$  (see Section 3.3). The asymptotic properties of (3.9) will be discussed in Section 4.

Specifically, as predictor of  $X_i$ , we propose

$$(3.10) \quad \hat{X}_i^{\mathcal{M}} = \frac{\sum_{j \neq i} \kappa(H^{-1}(\hat{\psi}(X_i) - \hat{\psi}(X_j))) \hat{X}_j}{\sum_{j \neq i} \kappa(H^{-1}(\hat{\psi}(X_i) - \hat{\psi}(X_j)))},$$

borrowing strength from local neighbors in the  $d$ -dimensional representation space. This can be seen as an alternative to representation (3.2), where we use the FPCs and borrow strength from the whole data set to estimate functional mean and eigenbasis. As before, we note that (3.10) is not necessarily in  $\mathcal{M}$ , but will be in a small neighborhood asymptotically and in compar-

ison with (3.2), (3.10) usually proves to be a much better predictor of  $X_i$  for functional manifold data as shown in the simulations and applications in Section 5. Asymptotic properties are discussed in Section 4.

Definition (2.3) suggests to estimate the manifold mean by

$$(3.11) \quad \hat{\mu}^{\mathcal{M}} = \frac{\sum_i \kappa(H^{-1}(\hat{\psi}(X_i) - \hat{\mu})) \hat{X}_i}{\sum_i \kappa(H^{-1}(\hat{\psi}(X_i) - \hat{\mu}))},$$

where  $\hat{\mu} = \frac{1}{n} \sum_i \hat{\psi}(X_i)$ . The FMC vectors  $\mathbf{e}_j$  defined in (2.8) are estimated by eigendecomposition of the sample covariance matrix of  $\hat{\psi}(X_i)$ , that is,  $\hat{\lambda}_j^{\mathcal{M}}$  and  $\hat{\mathbf{e}}_j$  are such that

$$(3.12) \quad \sum_{j=1}^d \hat{\lambda}_j^{\mathcal{M}} \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j^T = \frac{1}{n-1} \left\{ \sum_{i=1}^n \hat{\psi}(X_i) \hat{\psi}^T(X_i) - \frac{1}{n} \left( \sum_{j=1}^n \hat{\psi}(X_j) \right) \left( \sum_{j=1}^n \hat{\psi}(X_j) \right)^T \right\},$$

where the  $\hat{\lambda}_j^{\mathcal{M}}$  are ordered to be nonincreasing in  $j$ . From (2.9) and (3.9), we obtain estimates of the manifold modes as

$$(3.13) \quad \hat{X}_{j,\alpha}^{\mathcal{M}} = \frac{\sum_i \kappa(H^{-1}\{\hat{\psi}(X_i) - \hat{\mu} - \alpha(\hat{\lambda}_j^{\mathcal{M}})^{1/2} \hat{\mathbf{e}}_j\}) \hat{X}_i}{\sum_i \kappa(H^{-1}\{\hat{\psi}(X_i) - \hat{\mu} - \alpha(\hat{\lambda}_j^{\mathcal{M}})^{1/2} \hat{\mathbf{e}}_j\})},$$

where  $j = 1, \dots, d$  and  $\alpha \in \mathbb{R}$ .

**3.3. Selection of auxiliary parameters.** We use 10-fold cross-validation to simultaneously choose the step size  $\varepsilon$ , the truncation parameter  $\delta$ , and the smoothing bandwidth  $h$  (see Sections 3.1 and 3.2). The number of candidates for  $\varepsilon$  and  $\delta$  is kept small so that the cross-validation procedure runs reasonably fast. Candidates for the step size  $\varepsilon$  are the median distance of the 5th, the 8th and the 12th nearest neighbor; those for  $\delta$  are selected such that 0%, 2%, 5% and 10% of the data with the lowest local density estimates are penalized. Each of 10 subgroups of curves denoted by  $V_1, \dots, V_{10}$  is used as a validation set, one at a time, while the remaining data are used as training set.

In an initial step, we use the whole data set and a given  $\varepsilon, \delta$  to determine  $\hat{\psi}(X_i)$ , followed by estimation of  $X_i = \psi^{-1}(\vartheta_i)$  for  $X_i$  in the validation set, using (3.9) and assuming that only those  $\hat{X}_j$  in the training set are known. Denoting the value of the estimated trajectory  $X_i$ , evaluated at time  $t_{il}$ , by  $\hat{X}_{il}$ , the sum of squared prediction errors for the validation set  $V_k$  is  $\text{SSPE}_k = \sum_{i \in V_k} \sum_{l=1}^{n_i} (\hat{X}_{il} - Y_{il})^2$ , where  $Y_{il} = X_i(t_{il}) + \varepsilon_{il}$  is the observed value of trajectory  $X_i$  at time  $t_{ij}$ . The cross-validation choice is the minimizer of  $\text{MSPE}(\varepsilon, h, \delta) = \frac{\sum_{k=1}^{10} \text{SSPE}_k}{\sum_{i=1}^{10} n_i}$ .

Following [31], the intrinsic dimension  $d$  can be chosen by the  $1 - \beta$  fraction of distances explained (FDE), that is,

$$(3.14) \quad d = \min_p \left\{ p : \frac{\|\hat{D}^p - D\|_F}{\|D\|_F} < \beta \right\},$$

where we choose  $\beta = 0.05$  and  $D, \hat{D}^p$  are  $n$  by  $n$  distance matrixes with  $D_{ij} = \hat{d}_g(X_i, X_j)$  as in (3.8),  $\hat{D}_{ij}^p = \|\hat{\psi}^p(X_i) - \hat{\psi}^p(X_j)\|$  and where  $\hat{\psi}^p$  denotes the MDS solution (3.1) in  $\mathbb{R}^p$ , and  $\|\cdot\|_F$  is the matrix Frobenius norm,  $\|D\|_F = \{\sum_{i,j} D_{ij}^2\}^{1/2}$ . Note that  $\|\hat{D}^p - D\|_F$  is the square root of the minimized value of (3.1).

**4. Asymptotic properties.** We provide the specific convergence rate of  $\hat{X}_i^K$ , defined in (3.2), under assumptions (A1)–(A5) in the Appendix. Note that condition (A3) requires that the random functions are sampled at a dense design. Our starting point is that the manifold can be well identified at the sample points through ISOMAP, or alternatively, that the ISOMAP identified manifold may be viewed as the target. The difference between the target and the identified manifold from ISOMAP is quantified by a rate  $v_n$  that is assumed as given; if the target manifold corresponds to the manifold as identified at the sample points, we may set  $v_n = 0$ . The theoretical analysis aims to justify the new manifold representations that we propose, and for this it is essential to consider the behavior of the estimates across the entire function space. Therefore, our theoretical results demonstrate how to extend local behavior at the sample points to obtain global consistency of the proposed functional manifold representations.

As the convergence is for  $K = K_n \rightarrow \infty$  as  $n \rightarrow \infty$ , the rate of decline of the eigenvalues in (2.5) and also lower bounds on the spacing of consecutive eigenvalues, as postulated in (A4) are relevant, with a requirement of polynomially fast declining eigenvalues. Required smoothness and boundedness assumptions for  $X \in \mathcal{M}$  are as in (A5).

**PROPOSITION 3.** *Assume (A1)–(A5) in the Appendix, and define  $r_n = \max\{\frac{1}{\sqrt{nh_G^2}}, \frac{1}{\sqrt{nh_\mu}}, \frac{1}{\sqrt{nh_V}}\}$ . If there are infinitely many nonzero eigenvalues  $\lambda_k$  in (2.5), which are all of multiplicity one, then for sequences  $K = K_n \rightarrow \infty$ , subject to  $r_n K^{\alpha_2+1/2} \rightarrow 0$ , where  $\alpha_2$  is a constant such that  $\lambda_k - \lambda_{k+1} > C_2 k^{-\alpha_2}$  for some  $C_2 > 0$  and where  $K \leq K_0$  with  $K_0 = \min\{i : \lambda_i - \lambda_{i+1} \leq 2D_n\} - 1$  and  $D_n = \{\int_{\mathcal{T}^2} (\hat{G}(t, s) - G(t, s))^2 dt ds\}^{1/2}$  where  $G$  is defined in (2.5) and  $\hat{G}$  is defined after (3.2), it holds that*

$$(4.1) \quad \|\hat{X}_i^K - X_i\|_{L^2} = O_p(r_n K^{\alpha_2+1/2} + K^{-(\alpha_1-1)/2})$$

for  $\hat{X}_i^K$  defined in (3.2), where  $\alpha_1$  is such that  $\lambda_k < C_1 k^{-\alpha_1}$  for all  $k$  and some  $C_1 < \infty$ .

We note that under the assumptions,  $K_0 \rightarrow \infty$ . The first term on the r.h.s. of (4.1) is due to estimation error and the second term is due to truncation error. In the special case when there are only finitely many nonzero  $\lambda_k$  in (2.5), it can be shown that the rate in (4.1) simply becomes  $O_p(r_n)$ . Next we discuss the convergence of the estimates that appear in (3.12).

PROPOSITION 4. *Under (B1) and (B2) in the Appendix,*

$$(4.2) \quad \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p\left(v_n + \frac{1}{\sqrt{n}}\right),$$

where  $\boldsymbol{\mu}$  and  $\hat{\boldsymbol{\mu}}$  are defined in (2.3) and (3.11), and  $v_n = \sup_{i=1, \dots, n} \|\hat{\psi}(X_i) - \psi(X_i)\|$ . If the  $j$ th eigenvalue of  $\text{Cov}(\psi(X))$  is of multiplicity one, then

$$(4.3) \quad \|\hat{\mathbf{e}}_j - \mathbf{e}_j\| = O_p\left(v_n + \frac{1}{\sqrt{n}}\right),$$

$$(4.4) \quad |\hat{\lambda}_j^{\mathcal{M}} - \lambda_j^{\mathcal{M}}| = O_p\left(v_n + \frac{1}{\sqrt{n}}\right),$$

where  $\lambda_j^{\mathcal{M}}$ ,  $\mathbf{e}_j$ ,  $\hat{\mathbf{e}}_j$  and  $\hat{\lambda}_j^{\mathcal{M}}$  are defined in (2.8) and (3.12), respectively.

THEOREM 1. *Under (A1)–(A5), (B1), (B2) and (C1)–(C3) in the Appendix, assume that the density function  $f$  of  $\psi(X) \in \mathbb{R}^d$  satisfies  $f(\boldsymbol{\theta}) > 0$  for a specific  $\boldsymbol{\theta} = \psi(x)$  and that  $h > 0$  is selected such that  $h \rightarrow 0$ ,  $n^{-1}h^{-2(d+1)} \rightarrow 0$  and  $h^{-(d+1)}\text{Ev}_n \rightarrow 0$ . Then  $\hat{\psi}^{-1}(\boldsymbol{\theta})$  defined in (3.9), using  $\hat{X}_i = \hat{X}_i^K$ , is a consistent estimate of  $\psi^{-1}(\boldsymbol{\theta})$ . Specifically, defining  $T_\phi^K = \{\sum_{k>K} \xi_k^2\}^{1/2}$  where  $\xi_k = \int (X - \text{EX})\phi_k$  and the orthonormal basis  $\{\phi_k : k \geq 1\}$  is given in (2.5), and defining  $R_K(\boldsymbol{\theta}) = T_\phi^K(\psi^{-1}(\boldsymbol{\theta}))$ , where  $R_K(\boldsymbol{\theta}) \rightarrow 0$  as  $K = K_n \rightarrow \infty$ , it holds that*

$$(4.5) \quad \begin{aligned} & \|\hat{\psi}^{-1}(\boldsymbol{\theta}) - \psi^{-1}(\boldsymbol{\theta})\|_{L^2} \\ &= O_p\left(h^2 + \frac{1}{\sqrt{nh^d}} + \frac{v_n}{h} + R_K(\boldsymbol{\theta}) + K^{\alpha_2+1/2}r_n\right), \end{aligned}$$

where  $r_n$ ,  $\alpha_2$  and  $v_n$  are as in assumptions (A3), (A4) and (B1).

Note that  $R_K(\boldsymbol{\theta})$  corresponds to the truncation error for  $\psi^{-1}(\boldsymbol{\theta}) \in \mathcal{M}$ . The last term  $K^{\alpha_2+1/2}r_n$  is due to the estimation error as in Lemma 1. The middle term  $\frac{v_n}{h}$  reflects the estimation error of the weights, which is influenced by the scale of the bandwidth. The first part  $h^2 + \frac{1}{\sqrt{nh^d}}$  is the optimal rate when the  $X_i$  and  $\psi$  are known, reflecting an intrinsically  $d$ -dimensional smoothing problem. Related findings are discussed in [3].

For the manifold modes, we obtain the following corollary.

COROLLARY 1. *Under the conditions of Theorem 1, for a given  $\alpha \in \mathbb{R}$  and  $1 \leq j \leq d$ , assume that  $f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{1/2}\mathbf{e}_j) > 0$  and that  $h$  is chosen as in Theorem 1. Then the estimated manifold modes  $\hat{X}_{j,\alpha}^{\mathcal{M}}$  as in (3.13),*

substituting  $\hat{X}_i = \hat{X}_i^K$ , are consistent. Specifically,

$$(4.6) \quad \begin{aligned} & \|\hat{X}_{j,\alpha}^{\mathcal{M}} - X_{j,\alpha}^{\mathcal{M}}\|_{L^2} \\ &= O_p\left(h^2 + \frac{1}{\sqrt{nh^d}} + \frac{v_n}{h} + \frac{1}{\sqrt{nh}} + R_K + K^{\alpha_2+1/2}r_n\right), \end{aligned}$$

where  $R_K = T_\phi^K(X_{j,\alpha}^{\mathcal{M}})$ .

An immediate consequence of these results is that the manifold representation given in (2.10) provides a consistent representation of all random functions in the functional manifold. Proofs of all propositions, theorem and corollary can be found in the supplementary file [5].

## 5. Examples and simulation study.

5.1. *Functional manifolds and isometry.* To illustrate our methods and to discuss the impact of the critical isometry assumption, we consider the following three example functional manifolds:

- (i) A one-dimensional ( $d = 1$ ) functional manifold

$$\begin{aligned} \mathcal{M}_1 = & \left\{ X \in L^2([-4, 4]) : X(t) = \mu(h_\alpha(t)), \right. \\ & \left. h_\alpha(t) = \frac{8 \int_0^{t/8+0.5} s^\alpha (1-s) ds}{\int_0^1 s^\alpha (1-s) ds} - 4, \alpha > -1 \right\}, \end{aligned}$$

where  $\mu(t) = \frac{2}{\sqrt{\pi}} \exp\{-\frac{1}{2}(t+2)^2\} + \frac{1}{\sqrt{2\pi}} \exp\{-2(t-2)^2\}$ . This corresponds to random warping of a common shape function  $\mu$ , which has two peaks. The time warping function  $h_\alpha$  is generated from the cumulative Beta distribution family and  $\alpha$  is a random parameter,  $\alpha = \max(-1, Z)$ , where  $Z \sim N(0, 0.09)$ .

- (ii) A two-dimensional ( $d = 2$ ) functional manifold

$$\begin{aligned} \mathcal{M}_2 = & \left\{ X \in L^2([-4, 4]) : X(t) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left[-\frac{1}{2\alpha^2}(t-\beta)^2\right], \right. \\ & \left. \alpha > 0, \beta \in \mathbb{R} \right\}. \end{aligned}$$

This manifold is a collection of Gaussian densities, corresponding to a shift-scale family, where  $\alpha = \max(0, Z)$ ,  $Z \sim N(1, 0.04)$  and  $\beta \sim N(0, 1)$ .

- (iii) Another two-dimensional ( $d = 2$ ) functional manifold

$$\begin{aligned} \mathcal{M}_3 = & \left\{ X \in L^2([-4, 4]) : X(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(t-0.8-\alpha)^2\right\} \right. \\ & \left. + \frac{1}{\sqrt{\pi}} \exp\{-(t+0.8-\beta)^2\}, \alpha, \beta \in \mathbb{R} \right\}, \end{aligned}$$

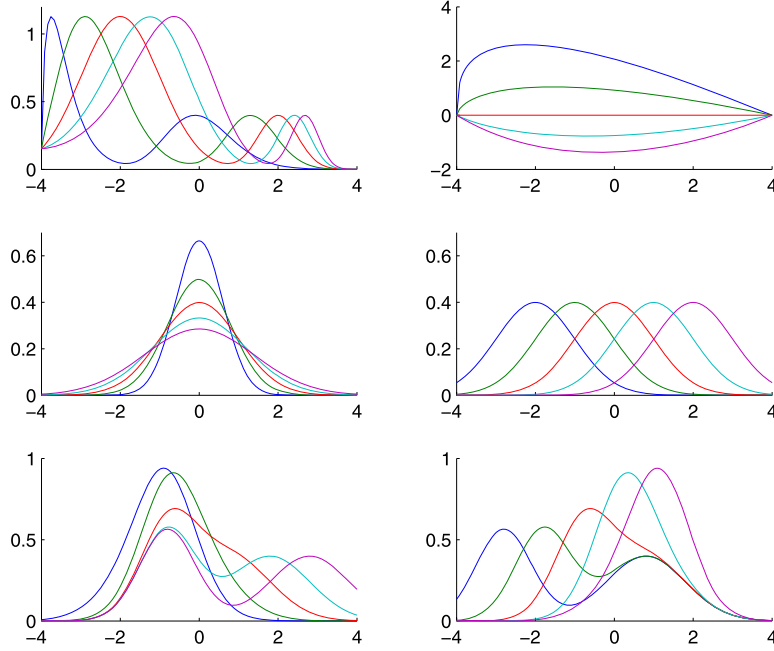


FIG. 1. Manifolds  $\mathcal{M}_1$ – $\mathcal{M}_3$ . Top left panel: functions on  $\mathcal{M}_1$  for  $\alpha = 0.6, 0.8, 1.0, 1.2, 1.4$ . Top right panel: corresponding identity-subtracted warping functions  $h_\alpha(t) - t$ . Middle left panel: functions on  $\mathcal{M}_2$  for  $\alpha = 0.4, 0.7, 1.0, 1.3, 1.6$  and  $\beta = 0$ . Middle right panel: functions on  $\mathcal{M}_2$  for  $\beta = 0.4, 0.7, 1.0, 1.3, 1.6$  and  $\alpha = 0$ . Bottom left panel: functions on  $\mathcal{M}_3$  for  $\alpha = -2, -1, 0, 1, 2$  and  $\beta = 0$ . Bottom right panel: functions on  $\mathcal{M}_3$  for  $\beta = -2, -1, 0, 1, 2$  and  $\alpha = 0$ .

a mixture of two peaks with randomly varying centers, where  $\alpha \sim \mathcal{N}(0, 1)$  and  $\beta \sim \mathcal{N}(0, 1)$ . Note that the two peaks will merge to a larger peak when their locations are close, so this set of functions has a randomly varying number of peaks.

Functional manifolds  $\mathcal{M}_1$ – $\mathcal{M}_3$  are illustrated in Figure 1. We note that  $\mathcal{M}_1$  is an isometric manifold and  $\mathcal{M}_2$  is approximately isometric, while  $\mathcal{M}_3$  is not isometric. This can be seen as follows. For functions  $X \in L^2$  on a differentiable isometric manifold with representation  $X = \psi^{-1}(\theta_1, \dots, \theta_d)$ , using the definition of isometry given after (2.1), the condition  $\int_{\theta_k^0}^{\theta_k^1} \left\| \frac{\partial X}{\partial \theta_k}(t) \right\|_{L^2} d\theta_k \equiv \theta_k^1 - \theta_k^0$  for  $k = 1, \dots, d$  and any  $\theta_k^0, \theta_k^1 \in \mathbb{R}$  is equivalent to isometry. Therefore, the existence of a parametrization of the map  $\psi$  for which the  $L^2$  norms of the partial derivatives of  $X$  with respect to the parameter components are constant is sufficient and necessary for  $\psi$  to be isometric. For one-dimensional manifolds such as  $\mathcal{M}_1$ , one can always find such a parametrization, as long as  $X$  is differentiable in the parameter and the derivative is  $L^2$  integrable in  $t$ .

TABLE 1  
*Fraction of distances explained (3.14) for isometric manifold fits with different dimension  $d$  (other parameters are optimized), for two signal-to-noise ratios  $R$*

Manifold	$R$	$d$				
		1	2	3	4	5
$\mathcal{M}_1$	0.1	0.998	0.999	0.999	0.999	0.999
	0.5	0.9778	0.993	0.995	0.995	0.996
$\mathcal{M}_2$	0.1	0.914	0.988	0.994	0.996	0.996
	0.5	0.902	0.971	0.974	0.978	0.980
$\mathcal{M}_3$	0.1	0.699	0.932	0.957	0.977	0.980
	0.5	0.639	0.906	0.948	0.955	0.958
Growth		0.947	0.972	0.980	0.985	0.988
Yeast		0.891	0.949	0.981	0.983	0.984
Mortality		0.878	0.954	0.973	0.980	0.982

For  $\mathcal{M}_2$ , such a parametrization does not exist, but since  $\|\frac{\partial X}{\partial \alpha}(t)\|_{L^2} = \frac{1}{\alpha}c_1$  and  $\|\frac{\partial X}{\partial \beta}(t)\|_{L^2} = \frac{1}{\alpha}c_2$  for constants  $c_1, c_2$  and as  $\alpha$  is chosen to remain very close to 1, the natural parametrization approximately satisfies the condition for isometry. In contrast to  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , the functional manifold  $\mathcal{M}_3$  is non-isometric and we include it as an example how the proposed methodology is faring when the key assumption of isometry is violated. As our considerations take place in a manifold learning framework, where the underlying manifold is unknown, an interesting aspect is to devise a data-based check to gauge the degree to which the isometry assumption can be expected to be satisfied. A natural metric for such a check is the fraction of distances explained (FDE), defined in (3.14). This criterion quantifies the percentage of geodesic distance that is preserved when fitting a  $d$ -dimensional isometric manifold to the data. For cases where the underlying manifold is actually nonisomorphic, the fitted manifold is an isometric approximation to the true underlying manifold, obtained by minimizing the stress function in the MDS algorithm.

An informal goodness-of-fit criterion for isometry is to require FDE to be larger than 95%, and choosing the manifold with the smallest dimension that satisfies this criterion. In Table 1, values for FDE obtained for the simulated data for manifolds  $\mathcal{M}_1$ – $\mathcal{M}_3$  under two signal-to-noise ratios  $R$  (defined in the following subsection) are reported, with dimension  $d$  ranging from 1 to 5. The well-known fact that the stress function declines when the dimension of the projection space is increased underlies the traditional MDS-Scree Plot [6] and is reflected by the observed increase in the values for FDE as dimension increases.

Applying the above check for isometry, we find that indeed the dimensions of the isometric manifold  $\mathcal{M}_1$  and the near-isometric manifold  $\mathcal{M}_2$  are

correctly selected, while the first two dimensions of the isometric manifold approximation to the nonisometric manifold  $\mathcal{M}_3$  are not sufficient. Thus, the nonisometric nature of  $\mathcal{M}_3$  means that the dimension of the underlying functional manifold cannot be correctly identified and instead the proposed algorithm will find a higher-dimensional isometric manifold to represent  $\mathcal{M}_3$ . The price to pay for a suitable isometric approximation is increased dimensionality, which in this example ends up larger than 2 for the approximating isometric manifold. We note that an approximating isometric manifold can always be found, since the linear and therefore intrinsically isometric manifold of infinite dimensionality that is spanned by the eigenfunction basis contains the random functions of the sample, according to the Karhunen–Loève theorem, and is always applicable.

While we can always find a near-isometric manifold of large enough dimensionality with the proposed algorithm, when the data lie on a lower-dimensional nonisometric manifold, these approximating isometric manifolds may not be efficient, since they do not provide the lowest-dimensional possible description of the data. Nevertheless, an approximating isometric nonlinear manifold obtained by the proposed approach often will present a much improved and lower-dimensional description when compared to the alternative of classical linear basis representation. This is exemplified by the functional nonisometric manifold  $\mathcal{M}_3$ , which in the following subsection is shown to be much better represented by an isometric manifold than by a linear basis. So the price that the isometry assumption exacts in nonisometric situations is that the proposed approach leads to a more or less suboptimal representation, which however will often be substantially lower-dimensional than an equally adequate linear representation. We conclude that even in nonisometric situations the proposed approach can often be expected to lead to improved representations of functional data.

**5.2. Simulation results.** We simulate functional data from manifolds  $\mathcal{M}_1$ – $\mathcal{M}_3$  as introduced in the previous subsection, aiming to study two questions. First, when the functional data lie on a manifold, whether it is isometric or not, does the proposed functional manifold approach lead to better (more parsimonious, better interpretable) representations of the data, compared to functional principal component analysis? Second, for noisy functional data that do not exactly lie on a manifold, how much improvement may one gain by adding the data-adaptive penalties implemented by P-ISOMAP, as described in Section 3.1?

For these simulations, the actual error-contaminated observations of the functional trajectories are generated as  $Y_{ij} = X_i(t_{ij}) + \varepsilon_{ij}$ ,  $\varepsilon_{ij} \sim N(0, \sigma^2)$  i.i.d.,  $i = 1, \dots, n$ ,  $j = 1, \dots, n_i$ , where  $n = 200$ ,  $t_{ij}$  equally spaced in  $[-4, 4]$  with 30 observations per trajectory, and the noise variance  $\sigma^2$  is such that the signal-to-noise ratio  $R$  is 0.1 or 0.5. We estimated manifold means  $\mu^{\mathcal{M}}$  (2.3),



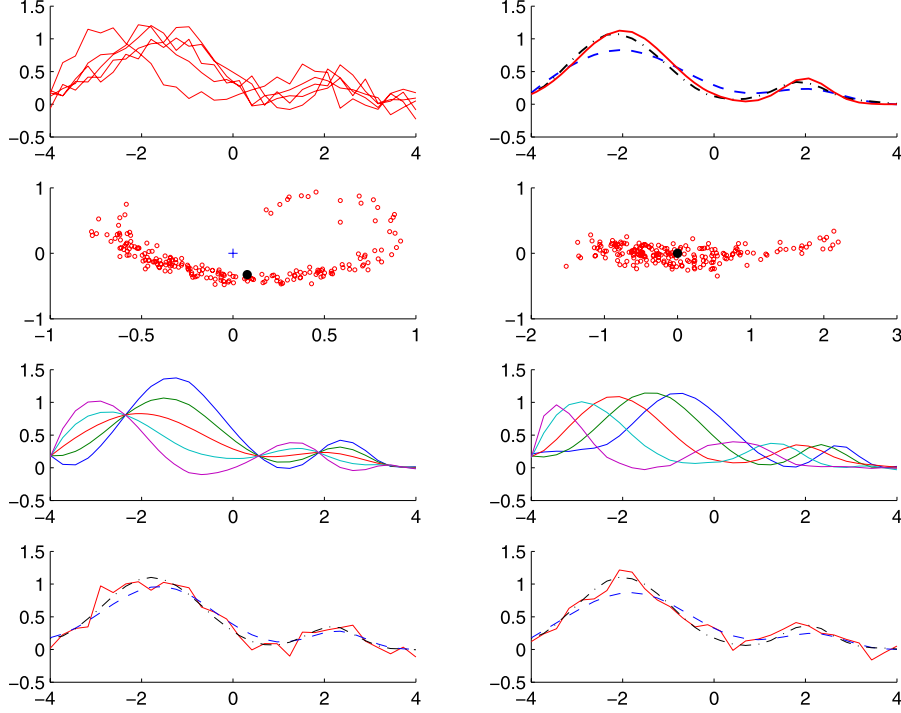


FIG. 2. Simulated data for manifold  $\mathcal{M}_1$ . Here and in the following figures, color descriptions refer to the online version of the paper. Top left panel: five randomly selected curves. Top right panel: common shape function (solid red, corresponds to target mean), estimated manifold mean  $\hat{\mu}^{\mathcal{M}}$  (3.11) (dash-dot black) and the  $L^2$  mean (dashed blue). Second row: scatter plot of second versus first functional principal component (left) and second versus first functional manifold component (right), where the bold black dot represents the manifold mean and the blue cross dot represents the  $L^2$  mean. Third row: estimates of principal component based mode  $X_{1,\alpha}$  (2.7) (left) and of manifold mode  $X_{1,\alpha}^{\mathcal{M}}$  (2.9) (right) of functional variation for  $\alpha = -2, -1, 0, 1, 2$ . Bottom row: two randomly selected curves (solid red), with the corresponding principal component based predictions  $\hat{X}_i^L$  (3.2) (dashed blue), and manifold based predictions  $\hat{X}_i^{\mathcal{M}}$  (3.10) (dash-dot black) for  $L = d = 2$ .

manifold modes of functional variation  $X_{j,\alpha}^{\mathcal{M}}$  (2.9) and obtained predictions  $\hat{X}_i^{\mathcal{M}}$  (3.10), which were compared with predictions obtained by functional principal component analysis.

Results for a simulation run are shown in Figures 2, 3 and 4 for manifolds  $\mathcal{M}_1$ – $\mathcal{M}_3$ , respectively. The estimated manifold means are seen to be close to the corresponding intrinsic means, that is, the common shape function for manifold  $\mathcal{M}_1$ , the standard Gaussian density for manifold  $\mathcal{M}_2$  and the curve with no time shifts ( $\alpha = \beta = 0$ ) for manifold  $\mathcal{M}_3$ . On the other hand, the cross-sectional means are seen to be far away from these intrinsic means and therefore clearly are not useful as measures of location for these sets of functions.

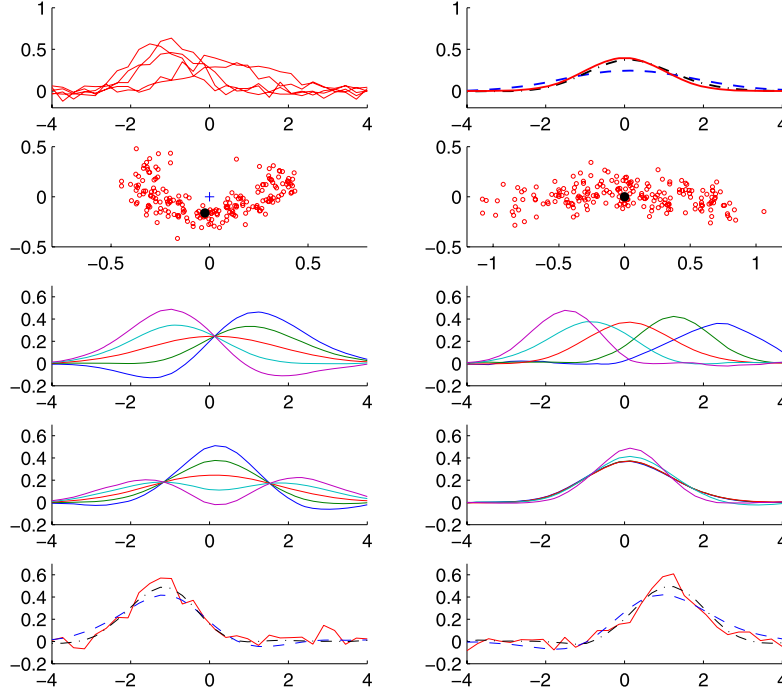


FIG. 3. Simulated data for manifold  $\mathcal{M}_2$ . Top left panel: five randomly selected curves. Top right panel: standard Gaussian density (solid red, corresponds to target mean), estimated manifold mean  $\hat{\mu}^{\mathcal{M}}$  (3.11) (dash-dot black) and the  $L^2$  mean (dashed blue). Second row: scatter plot of second versus first FPC (left) and second versus first FMC (right), where the bold black dot represents the manifold mean and the blue cross represents the  $L^2$  mean. Third row: estimates of principal component based mode  $X_{1,\alpha}$  (2.7) (left) and of manifold mode  $X_{1,\alpha}^{\mathcal{M}}$  (2.9) (right) of functional variation for  $\alpha = -2, -1, 0, 1, 2$ . Fourth row: estimates of  $X_{2,\alpha}$  (left) and of  $X_{2,\alpha}^{\mathcal{M}}$  (right) for  $\alpha = -2, -1, 0, 1, 2$ . Bottom row: two randomly selected curves (solid red), with the corresponding principal component based predictions  $\hat{X}_i^L$  (3.2) (dash blue), and manifold based predictions  $\hat{X}_i^{\mathcal{M}}$  (3.10) (dash-dot black) for  $L = d = 3$ .

The scatter plots of second versus first FPC indicate “horseshoe” shapes for manifolds  $\mathcal{M}_1$  and  $\mathcal{M}_2$ . This diagnostic indicates that a functional manifold approach may be called for. We find that the location of the cross-sectional mean (at the origin, due to the zero expectation property of FPCs) typically lies in a relatively sparse region of the data in these scatter plots, while the manifold mean falls into a much denser area, which is another diagnostic feature pointing to an underlying manifold. Complex two-dimensional surface curvature is observed for manifold  $\mathcal{M}_3$ . Comparing with Figure 1, we find that the manifold modes represent the inherent components of functional variation present in the data quite well, while the established principal component based modes are not informative in describing the functional

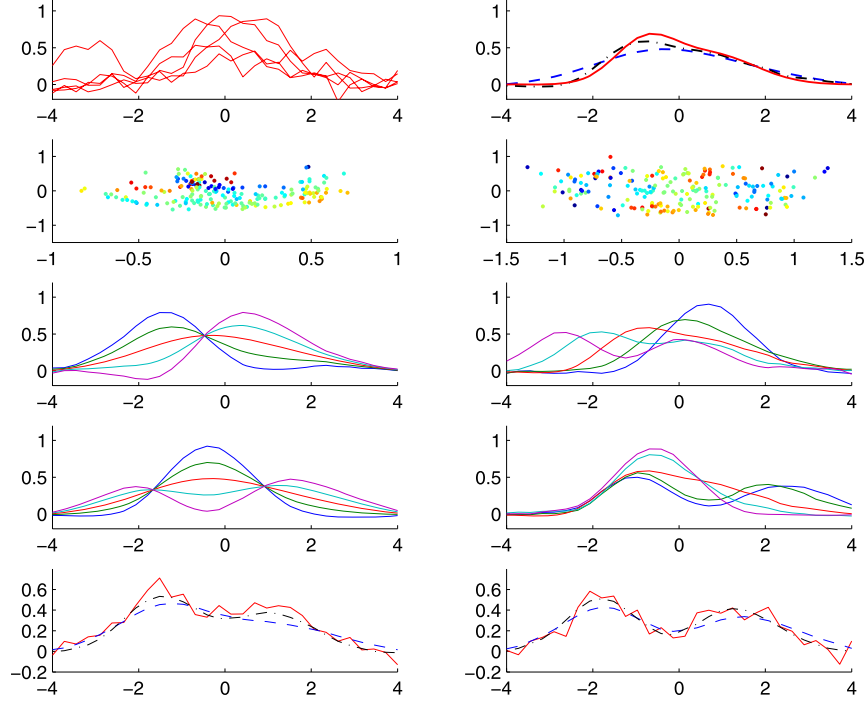


FIG. 4. Simulated data for manifold  $\mathcal{M}_3$ . Top left panel: five randomly selected curves. Top right panel: curve with no time shifts (solid red, corresponds to target mean), estimated manifold mean  $\hat{\mu}^{\mathcal{M}}$  (3.11) (dash-dot black) and the  $L^2$  mean (dash blue). Second row: contour scatter plot of second versus first FPC (left) and second versus first FMC (right), with the colors scaled from the third FPC or FMC. Third row: estimates of principal component based mode  $X_{1,\alpha}$  (2.7) (left) and of manifold mode  $X_{1,\alpha}^{\mathcal{M}}$  (2.9) (right) of functional variation for  $\alpha = -2, -1, 0, 1, 2$ . Fourth row: estimates of  $X_{2,\alpha}$  (left) and of  $X_{2,\alpha}^{\mathcal{M}}$  (right) for  $\alpha = -2, -1, 0, 1, 2$ . Bottom row: two randomly selected curves (solid red), with the corresponding principal component based predictions  $\hat{X}_i^L$  (3.2) (dash blue), and manifold based predictions  $\hat{X}_i^{\mathcal{M}}$  (3.10) (dash-dot black) for  $L = d = 3$ .

variation. It is also obvious that the proposed predictions for individual trajectories  $X_i$  are more accurate in capturing amplitudes and locations of peaks.

Leave-one-out predictions of the  $X_i$  are calculated using both functional principal components (3.2), resulting in  $\hat{X}_i^L$ , as well as the proposed new estimates  $\hat{X}_i^{\mathcal{M}}$  (3.10). For  $\hat{X}_i^L$ , we estimate the FPCs (2.6) of  $X_i$  using all data and then leave  $X_i$  out to obtain  $\hat{\mu}$  and  $\hat{\phi}_k$ ; for  $\hat{X}_i^{\mathcal{M}}$ , we estimate  $\hat{\psi}(X_i)$  using all data and then leave  $X_i$  out in the local averaging step. Starting with  $L = 1, d = 1$ , we increase  $L$  and  $d$  successively, obtaining the mean squared prediction errors  $\text{MSPE} = \frac{1}{200} \sum_{i=1}^{200} \|X_i - \hat{X}_i\|_{L^2}^2$ , where  $\hat{X}_i = \hat{X}_i^L$  or  $\hat{X}_i^{\mathcal{M}}$ , for  $1 \leq d = L \leq 5$ .

TABLE 2  
Mean squared prediction errors and relative squared prediction errors for  $\mathcal{M}_1$ – $\mathcal{M}_3$

		$R$	Method	MSPE with $L$ or $d$					RSPE with $L$ or $d$ (%)				
				1	2	3	4	5	1	2	3	4	5
$\mathcal{M}_1$	0.1	$\hat{X}_i^L$	0.159	0.034	0.025	0.021	0.021	41	10	6	6	6	
		$\hat{X}_i^{\mathcal{M}}$	0.027	0.015	0.015	0.014	0.015	7	4	4	4	4	
	0.5	$\hat{X}_i^L$	0.173	0.061	0.057	0.058	0.058	45	16	15	15	15	
		$\hat{X}_i^{\mathcal{M}}$	0.090	0.046	0.046	0.049	0.053	23	12	12	13	14	
$\mathcal{M}_2$	0.1	$\hat{X}_i^L$	0.054	0.022	0.013	0.008	0.007	44	17	10	7	6	
		$\hat{X}_i^{\mathcal{M}}$	0.022	0.009	0.007	0.006	0.006	18	8	5	5	5	
	0.5	$\hat{X}_i^L$	0.055	0.025	0.019	0.018	0.018	45	20	16	14	14	
		$\hat{X}_i^{\mathcal{M}}$	0.030	0.017	0.015	0.014	0.014	25	13	12	12	12	
$\mathcal{M}_3$	0.1	$\hat{X}_i^L$	0.148	0.059	0.031	0.023	0.020	59	23	13	9	7	
		$\hat{X}_i^{\mathcal{M}}$	0.088	0.025	0.020	0.020	0.019	35	10	8	8	8	
	0.5	$\hat{X}_i^L$	0.154	0.071	0.053	0.048	0.048	61	28	21	19	19	
		$\hat{X}_i^{\mathcal{M}}$	0.124	0.059	0.047	0.045	0.044	49	24	19	18	18	

The simulation results for manifolds  $\mathcal{M}_1$ – $\mathcal{M}_3$  are shown in Table 2. Generally, the MSPE is reduced by 20% over the established linear method when using the manifold approach; this improvement exceeds 50% when  $L$  and  $d$  are small. Another metric of interest is the relative squared prediction error of the model over the squared error when using the mean as predictor,  $\text{RSPE} = \frac{\sum_{i=1}^{200} \|X_i - \hat{X}_i\|_{L^2}^2}{\sum_{i=1}^{200} \|X_i - \bar{X}\|_{L^2}^2}$ , where  $\bar{X} = \frac{1}{200} \sum_{i=1}^{200} X_i$ , which can be interpreted as fraction of variance that is left unexplained. In all three simulated manifolds, RSPE is found to be much larger for the functional principal component representations, when the same number of components is used. This is because in the inefficient linear representation higher order functional principal components carry substantial variation.

To quantify the efficiency of the data-adaptive penalties in the proposed P-ISOMAP procedure, we also calculated the MSPE using the unmodified ISOMAP. Parameters for ISOMAP were selected analogously to the description in Section 3.3 by cross-validation. Since the most important comparison is for the case where  $d$  equals the intrinsic dimension, that is, 1 for  $\mathcal{M}_1$  and 2 for  $\mathcal{M}_2$  and  $\mathcal{M}_3$ , we calculated the ratio of the MSPE of P-ISOMAP over the MSPE of ISOMAP for these situations (Table 3). As anticipated, P-ISOMAP indeed exhibits increasing benefits for smaller signal-to-noise ratios.

The influence of the selection of the step size parameter  $\varepsilon$  in P-ISOMAP, defined in (3.7), on mean squared prediction errors is demonstrated in Ta-

TABLE 3  
*Mean squared prediction error  
ratios for P-ISOMAP over  
ISOMAP*

$R$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
0.1	0.9676	0.9679	0.9402
0.5	0.8121	0.8879	0.8302

ble 4. Here  $d$  is fixed as the intrinsic dimension (1 for  $\mathcal{M}_1$  and 2 for  $\mathcal{M}_2$ ,  $\mathcal{M}_3$ ), while  $\delta$  and  $h$  are optimized by cross-validation for each  $\varepsilon$ . We then select  $\varepsilon$  from the median distances of the 3rd, 5th, 8th, 12th and 16th nearest points calculated over all sample data. From the results in the table, one finds that the results are not strongly sensitive to the selection of  $\varepsilon$ , as long as it is in medium range. A good overall choice is median distance of 8th nearest neighbors. When  $\varepsilon$  is chosen very small, some sample points that are not situated close to other sample points may become separated from the other data, or disconnected subgroups in the data may emerge, which renders the MSPE for small  $\varepsilon$  inaccurate. In practice, we therefore impose a lower bound on  $\varepsilon$  to ensure that the fraction of data that are not connected to other points when connecting through  $\varepsilon$ -neighborhoods stays below 5%.

## 6. Applications.

6.1. *Berkeley growth study.* In growth studies, one often observes phase variation in the trajectories. Some subjects reach certain growth stages (such as puberty in human growth) earlier than others. This leads to difficulties for the parsimonious modeling of growth patterns with linear methods, and more generally for methods that are based on  $L^2$  distance between trajectories. Accordingly, cross-sectional mean estimation tends to fail in repre-

TABLE 4  
*Mean squared prediction errors using different  $\varepsilon$  for P-ISOMAP*

Manifold	$R$	$\varepsilon$				
		3	5	8	12	16
$\mathcal{M}_1$	0.1	0.029	0.027	0.031	0.029	0.033
	0.5	0.116	0.102	0.090	0.129	0.135
$\mathcal{M}_2$	0.1	0.008	0.010	0.009	0.010	0.010
	0.5	0.020	0.018	0.018	0.017	0.017
$\mathcal{M}_3$	0.1	0.029	0.040	0.025	0.27	0.033
	0.5	0.052	0.059	0.065	0.059	0.066

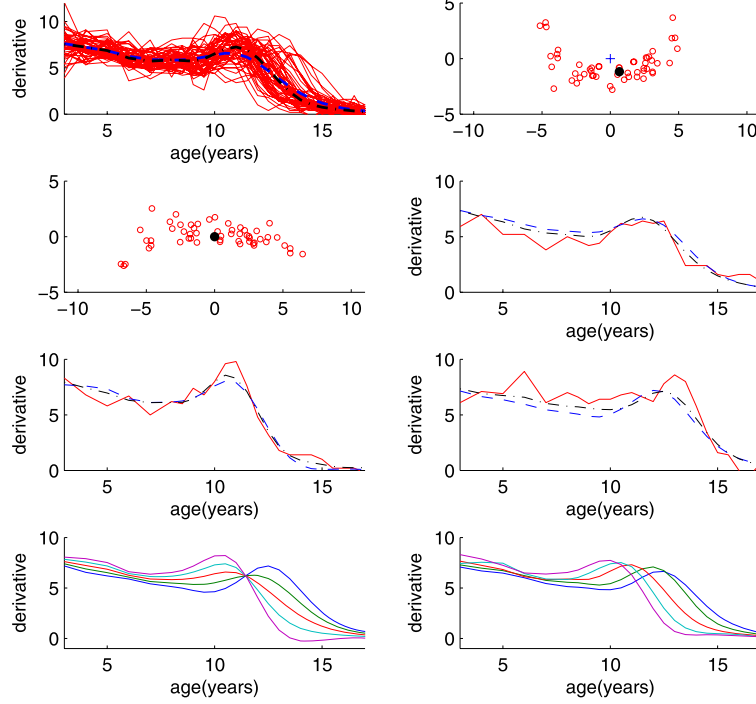


FIG. 5. Berkeley growth data for girls. Top left panel: derivatives with the cross-sectional mean (dash blue) and estimated manifold mean  $\hat{\mu}^M$  (3.11) (dash-dot black). Top right panel: scatter plot of second versus first FPC, where the bold black dot represents the manifold mean and the blue cross represents the cross-sectional mean. Second row left panel: scatter plot of second versus first FMC, where the bold black dot represents the manifold mean. Second row right panel and third row panels: three randomly selected curves (solid red), with the corresponding principal component based predictions  $\hat{X}_i^L$  (3.2) (dash blue), and manifold based predictions  $\hat{X}_i^M$  (3.10) (dash-dot black) for  $L = d = 2$ . Bottom panels: estimates of principal component based mode  $X_{1,\alpha}$  (2.7) (left) and of manifold mode  $X_{1,\alpha}^M$  (2.9) (right) of functional variation for  $\alpha = -2, -1, 0, 1, 2$ .

senting important growth features adequately [13, 22]. Since phase variation introduces nonlinear features in functional data, it is of interest to determine whether the analysis of growth data may benefit from the manifold approach.

We apply the manifold approach to the Berkeley growth data for females [32]. The data contain height measurements for 54 girls, with 31 measurements taken between the ages of 1 and 18 years. Interest usually focuses on growth velocity [11], which we obtain by smoothing the first-order difference quotients of the curves. The resulting growth velocity curves are shown in the top left panel of Figure 5, together with the cross-sectional mean and the estimated manifold mean  $\hat{\mu}^M$  (3.11). Similarly to Figures 2–4, the de-

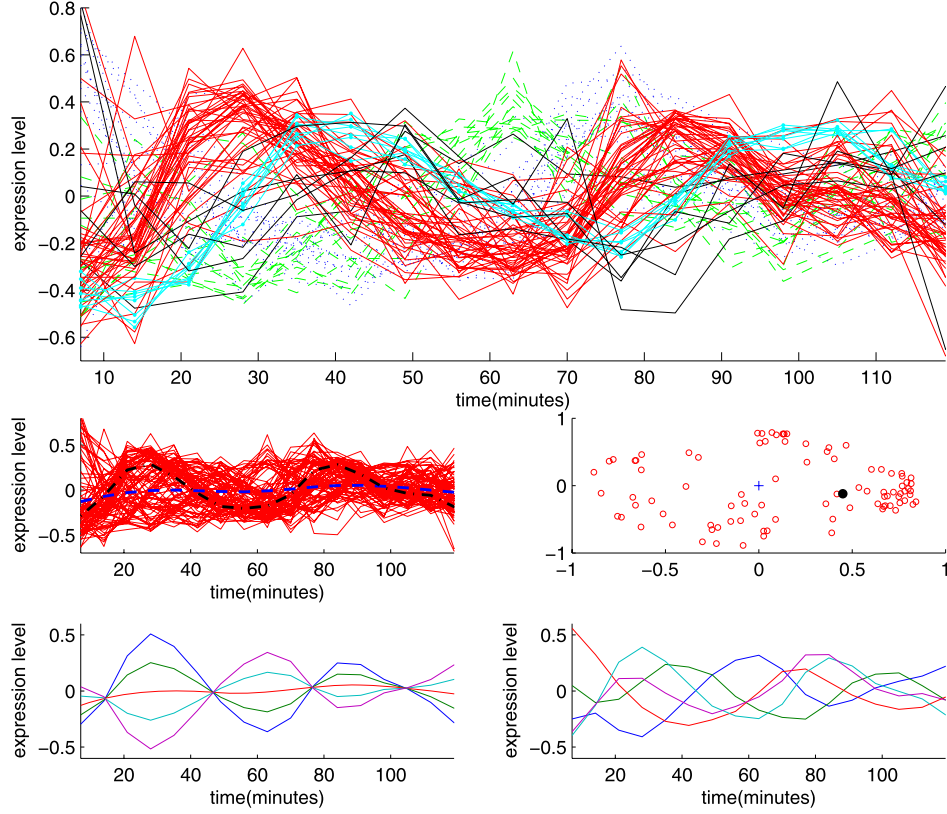


FIG. 6. Yeast cell cycle gene expression data. Top panel: all trajectories in different colors according to cluster membership: G1 (solid red), S (dash-dot cyan), G2/M (dash green), M/G1 (dotted blue) and S/G2 (solid black). Middle left panel: estimated manifold mean  $\hat{\mu}^M$  (3.11) (dash-dot black) and cross-sectional mean (dash blue). Middle right panel: scatter plot of second versus first FPC, where the blue cross indicates cross-sectional mean and the bold black dot indicates manifold mean. Bottom panels: estimates of principal component based mode  $X_{1,\alpha}$  (2.7) (left) and of manifold mode  $X_{1,\alpha}^M$  (2.9) (right) of functional variation for  $\alpha = -2, -1, 0, 1, 2$ .

descriptions of Figures 5–7 refer to the color online versions. The location of the cross-sectional mean, which falls at (0,0), and the location of the estimated manifold mean are indicated in the scatter plot of second versus first FPC (top right panel), which displays the “horseshoe” pattern described above. This, and the fact that the cross-sectional mean is away from the main data cloud, point to inherent nonlinearity in these data.

Mean squared prediction errors (MSPE) and relative squared prediction errors (RSPE) for the leave-one-out predictions of  $X_i$ , as described in Section 5, are listed in Table 5. The fractions of distance explained (FDE), de-

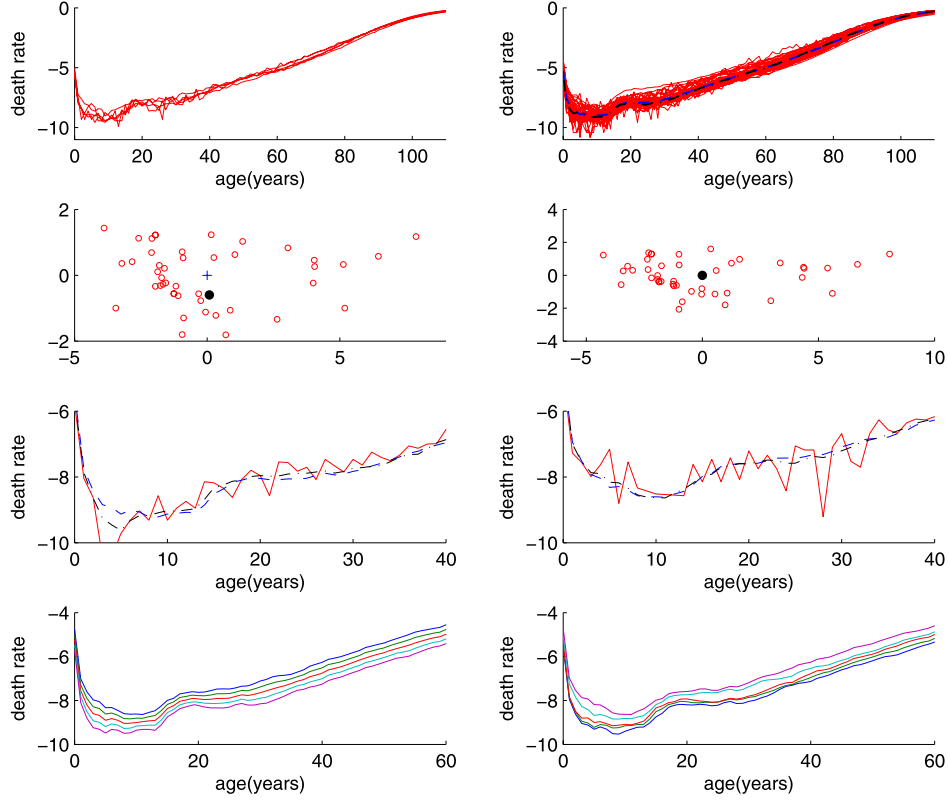


FIG. 7. Human mortality data. Top left panel: death rates for five randomly selected countries. Top right panel: estimates of cross-sectional mean (dash blue) and manifold mean  $\hat{\mu}^M$  (3.3) (dash-dot black). Second row: scatter plots of second versus first FPC (left) and second versus first FMC (right), where the blue cross indicates the cross-sectional mean and the bold black dot indicates the manifold mean. Third row: two randomly selected curves (solid red), with the corresponding principal component based predictions  $\hat{X}_i^L$  (3.2) (dash blue), and manifold based predictions  $\hat{X}_i^M$  (3.10) (dash-dot black) for  $L = d = 3$ . Bottom panels: estimates of principal component based mode  $X_{1,\alpha}$  (2.7) (left) and of manifold mode  $X_{1,\alpha}^M$  (2.9) (right) of functional variation for  $\alpha = -2, -1, 0, 1, 2$ .

fined in (3.14), for different dimensions  $d$  are shown in Table 1. The MSPE of  $\hat{X}_j^L$  is minimized at  $L = 5$ , with  $L = 2$  already a quite good choice.

We find that  $\hat{X}_j^M$  consistently improves upon  $\hat{X}_j^L$ , the fit obtained from functional principal components. Note that we used the preliminary estimator  $\hat{X}_i^K$  in (3.10) with  $K = 4$ , applying criterion (3.5). The FDE criterion indicates that these data can be well described by a one-dimensional manifold. The middle three panels of Figure 5 include three randomly selected curves, along with the predictions  $\hat{X}_i^L$  and  $\hat{X}_i^M$  using  $L = d = 2$ . The two bottom panels of Figure 5 illustrate the comparison of estimated manifold



TABLE 5  
Mean squared prediction errors and relative squared prediction errors for growth, yeast  
and mortality data

Data	Method	MSPE with $L$ or $d$					RSPE with $L$ or $d$ (%)				
		1	2	3	4	5	1	2	3	4	5
Growth	$\hat{X}_i^L$	17.1	12.9	13.8	13.7	12.6	62	47	50	50	46
	$\hat{X}_i^M$	10.7	9.46	9.06	9.21	9.08	39	34	33	33	33
Yeast	$\hat{X}_i^L$	0.639	0.382	0.257	0.205	0.203	67	40	27	22	21
	$\hat{X}_i^M$	0.468	0.278	0.231	0.210	0.206	49	29	24	22	22
Mortality	$\hat{X}_i^L$	7.38	6.34	5.44	5.48	5.21	54	47	40	40	38
	$\hat{X}_i^M$	6.77	5.64	5.40	5.26	4.98	50	41	40	39	37

modes of functional variation with the principal component based modes. The manifold modes are clearly more useful and adequately reflect the time-warping feature of these data. The first manifold mode specifically suggests that for girls, a puberty growth peak at a late age, especially after age 12, tends to have a smaller amplitude; this is in line with auxological knowledge. Overall, the manifold mode is seen to provide a clearer and much more adequate description of the longitudinal dynamics of these data.

6.2. *Yeast cell cycle gene expression.* Temporal expression curves for yeast cell cycle related genes were obtained by [29]. There are 6,178 genes in total, where each gene expression time-course consists of 18 data points, measured every 7 minutes between 0 and 119 minutes. Groups of genes are thought to be coexpressed coherently across different time periods, according to the role played by the genes in the time progression of the cell cycle. The dynamics of the gene expression levels are complex. Temporal regularization of gene expression is a characteristic of gene function, suggesting models that incorporate time-warping [24, 30].

The data we study consist of 90 genes that have been identified by biological methods [29]. Of these genes, 44 are thought to be related to G1 phase regulation of the yeast cell cycle and 46 to non-G1 phase regulation (S, S/G2, G2/M and M/G1 phases). Time courses of gene expression (top panel of Figure 6) for these clusters reveal two peaks for the G1 (solid red) and S (dash-dot cyan) groups, and one peak for G2/M (dash green) and M/G1 (dotted blue) groups, while the trajectories for the S/G2 (solid black) group are highly variable with no obvious peak.

The proposed manifold analysis was applied to this set of 90 genes. The estimated manifold mean  $\hat{\mu}^M$  (3.11) (middle left panel of Figure 6) is seen to fall within the G1 group (solid red in the top panel). In contrast, the cross-sectional mean is almost flat and does not reflect useful information about

these data. We also calculated the MSPE (Table 5) of  $\hat{X}_i^L$  (3.2) and  $\hat{X}_i^M$  (3.10), using preliminary estimators  $\hat{X}_i^K$  with  $K = 4$  in (3.10). The manifold-based predictions are seen to be much better for  $d = 1$  and 2, while they become more similar in performance to  $\hat{X}_i^L$  when  $d$  increases.

In the two bottom panels of Figure 6, we display the estimated manifold mode (right) and the principal component based mode of functional variation (left). The latter is found to be deceptive, as it indicates amplitude variation around a few fixed “knots,” while the first manifold mode clearly illustrates the actual temporal variation in the data, which is mainly caused by phase shifts. Each of the five groups, except the S/G2 group (solid black), is well represented by the variation across this manifold mode.

**6.3. Human mortality across countries.** The death rates derived from current life table cohorts for 44 countries in the year 2000, recorded for each age ranging from 0 to 110, have been collected and are as described in <http://www.lifetable.de/>. Death rates are widely used for descriptive and analytical purposes in public health, and cross-country comparisons are of particular interest here.

We view log-transformed annual death rates as noisy measurements of underlying smooth trajectories. Five sample trajectories are shown in the top left panel of Figure 7. The mortality trajectories are densely sampled, but the annual rates are quite noisy. We presmoothed this data, following (3.6). The resulting MSPEs for  $\hat{X}_i^M$  (3.10) and  $\hat{X}_i^L$  (3.2) are in Table 5. Manifold-based prediction is seen to perform better than linear principal component based prediction, regardless of the choice of dimension. This is also illustrated by the panels in the third row of Figure 7, where predicted trajectories are obtained for  $L = d = 3$ . For these data, the estimated manifold mean  $\hat{\mu}^M$  (3.11) does not differ dramatically from the cross-sectional mean (top right panel and second row left panel). However, the first manifold mode of variation (bottom right panel) indicates that countries with overall lower death rates, or more specifically, with death rates below the mean curve (solid red), exhibit less variation than those with death rates above the mean, especially for ages from 0 to 40. This finding is in line with the skewness that is apparent in the scatter plots, but is not seen in the principal component based mode (bottom left panel). The observed gains in prediction error for the manifold approach provide evidence that substantial nonlinearity is present in these data.

**7. Discussion.** While the proposed functional manifold implementations were running relatively fast on a linux server, observing that the computational complexity of classical MDS is of the order  $O(n^3)$ , computational difficulties may arise for truly large sample sizes  $n$ . In such situations, one might consider to base the proposed methods on landmark MDS [7], where one employs landmarks to significantly reduce the computational complexity.

The proposed method relies on two major assumptions: the isometry of the underlying functional manifold and that the target manifold is close or identical to the manifold identified by ISOMAP at the sample points. As for the isometry assumption, even if it is violated, the proposed method proves to be beneficial, as it often will provide for a much sparser representation of functional data in comparison with linear methods in cases where the underlying manifold is nonlinear, even if this manifold is not isometric. This is discussed in detail in Section 5.1 and borne out by simulations. As for the closeness of the ISOMAP solution to the true manifold at the sample points, this assumption and its underlying justification pertains to ISOMAP for vector data as proposed in [31].

Starting from the simplifying assumption that the ISOMAP identified manifold and the target manifold are essentially identical at the sample points, we proceed to extend the estimation of the manifold function to the entire space of interest. We note that such simplifying assumptions are often beneficial when deploying complex statistical methodology, as even when the assumptions are not completely satisfied, the resulting methodology may turn out to be more efficient than existing methods.

Overall, we find that the proposed manifold mean and manifold modes of functional variation provide useful representations that are competitive with and often superior over classical linear representations for functional data. The proposed functional manifold representations thus complement the established linear representations, notably the Karhunen–Loève representation, and in many instances provide more efficient models with better interpretations.

## APPENDIX: ASSUMPTIONS

(A1) The bandwidths  $h_\mu$ ,  $h_v$ ,  $h_G$  for estimating  $\mu(t)$ ,  $\sigma^2$ ,  $G(t, s)$  in Section 3.1 satisfy:  $h_\mu \rightarrow 0$ ,  $nh_\mu^4 \rightarrow \infty$  and  $nh_\mu^6 < \infty$ ;  $h_G \rightarrow 0$ ,  $nh_G^6 \rightarrow \infty$  and  $nh_G^8 < \infty$ ;  $h_v \rightarrow 0$ ,  $nh_v^4 \rightarrow \infty$  and  $nh_v^6 < \infty$ .

(A2) The smoothing kernels  $\kappa_\mu$  for the mean function  $\mu$  and  $\kappa_G$  for the covariance function  $G$  in Section 3.1 are absolutely integrable, that is,  $\int |\kappa_\mu(t)| dt < \infty$  and  $\iint |\kappa_G(t, s)| dt ds < \infty$ .

(A3) For  $\tau_{ij} = t_{ij} - t_{i,j-1}$  and  $\tau^* = \max_{i,j} \tau_{ij}$ , it holds that  $\tau^* = O_p(r_n^2)$ , where  $r_n = \max\{\frac{1}{\sqrt{nh_G^2}}, \frac{1}{\sqrt{nh_\mu}}, \frac{1}{\sqrt{nh_v}}\}$ .

(A4) The eigenvalues of the covariance function  $G(t, s)$  satisfy  $\lambda_k < C_1 k^{-\alpha_1}$  for some constants  $C_1 < \infty$ ,  $\alpha_1 > 1$ , and if  $\lambda_k > 0$ , then  $\lambda_k - \lambda_{k+1} > C_2 k^{-\alpha_2}$  for some constants  $C_2 > 0$  and  $\alpha_2 > 0$ .

(A5) For any  $X \in \mathcal{M}$ ,  $X$  is differentiable and  $\|X\|_\infty = O_p(1)$ ,  $\|X'\|_\infty = O_p(1)$ . The covariance function  $G(t, s)$  is twice differentiable in both  $t$  and  $s$ , and  $\sup_{t,s \in \mathcal{T}} |G(t, s)| < C_3$ ,  $\sup_{t,s \in \mathcal{T}} |\frac{\partial^2 G(t, s)}{\partial t \partial s}| < C_4$  for some constants  $C_3$ ,  $C_4 < \infty$ .

(B1) The estimates  $\hat{\psi}$  of  $\psi$  converge uniformly on the sample space, that is,  $Ev_n \rightarrow 0$  for  $v_n = \sup_{i=1,\dots,n} \|\hat{\psi}(X_i) - \psi(X_i)\|$ .

(B2) Each component of the  $d$ -vector  $\psi(X)$  has a finite fourth moment, and its covariance matrix is positive definite.

(C1) The  $d$ -vector  $\psi(X)$  admits a density function  $f$ , which is twice differentiable with continuous partial derivatives and uniformly bounded Hessian matrix.

(C2) The  $d$ -dimensional nonnegative kernel  $\kappa$  satisfies  $\int \kappa(\mathbf{u}) d\mathbf{u} = 1$ ,  $\kappa(\mathbf{u}) = \kappa(-\mathbf{u})$ ,  $\det(\int \kappa(\mathbf{u}) \mathbf{u} \mathbf{u}^T d\mathbf{u}) < \infty$ ,  $\int \kappa^2(\mathbf{u}) d\mathbf{u} < \infty$ , and is Lipschitz continuous with compact support,  $\{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| \leq 1\}$ .

(C3) The map  $\psi^{-1} : \mathbb{R}^d \rightarrow L^2$  is twice Fréchet differentiable, that is, there exist bounded linear operators  $A_{\mathbf{u}}^1 : \mathbb{R}^d \rightarrow L^2$ ,  $A_{\mathbf{u}}^2 : \mathbb{R}^d \times \mathbb{R}^d \rightarrow L^2$  such that

$$\lim_{\mathbf{u}_1 \rightarrow \mathbf{0}} \frac{\|\psi^{-1}(\mathbf{u} + \mathbf{u}_1) - \psi^{-1}(\mathbf{u}) - A_{\mathbf{u}}^1(\mathbf{u}_1)\|_{L^2}}{\|\mathbf{u}_1\|} = 0,$$

$$\lim_{\mathbf{u}_2 \rightarrow \mathbf{0}} \frac{\|A_{\mathbf{u}+\mathbf{u}_2}^1(\mathbf{u}_1) - A_{\mathbf{u}+\mathbf{u}_2}^1(\mathbf{u}_1) - A_{\mathbf{u}}^2(\mathbf{u}_1, \mathbf{u}_2)\|_{L^2}}{\|\mathbf{u}_2\|} = 0$$

for all  $\mathbf{u}, \mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^d$ . In addition,  $\frac{\|A_{\mathbf{u}}^2(\mathbf{u}_1, \mathbf{u}_2)\|_{L^2}}{\|\mathbf{u}_1\| \cdot \|\mathbf{u}_2\|}$  is continuous and uniformly bounded w.r.t.  $\mathbf{u}$ .

**Acknowledgments.** We are grateful for helpful comments on previous versions of this paper from two referees and an Associate Editor.

## SUPPLEMENTARY MATERIAL

**Supplement to “Nonlinear manifold representations for functional data”** (DOI: [10.1214/11-AOS936SUPP](https://doi.org/10.1214/11-AOS936SUPP); .pdf). An online supplementary file contains the detailed proofs for Propositions 1–4, Theorem 1 and Corollary 1. These proofs make use of material in references [15, 20, 25, 26].

## REFERENCES

- [1] ASH, R. B. and GARDNER, M. F. (1975). *Topics in Stochastic Processes. Probability and Mathematical Statistics* **27**. Academic Press, New York. [MR0448463](#)
- [2] BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.
- [3] BICKEL, P. J. and LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems. IMS Lecture Notes Monogr. Ser.* **54** 177–186. IMS, Beachwood, OH. [MR2459188](#)
- [4] CASTRO, P. E., LAWTON, W. H. and SYLVESTRE, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28** 329–337.
- [5] CHEN, D. and MÜLLER, H. G. (2011). Supplement to “Nonlinear manifold representations for functional data.” [DOI:10.1214/11-AOS936SUPP](https://doi.org/10.1214/11-AOS936SUPP).

- [6] COX, T. F. and COX, M. A. A. (2001). *Multidimensional Scaling*. Chapman and Hall, London.
- [7] DE SILVA, V. and TENENBAUM, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. *Adv. Neural Inf. Process. Syst.* **15** 721–728.
- [8] DO CARMO, M. P. (1992). *Riemannian Geometry*. Birkhäuser, Boston, MA. [MR1138207](#)
- [9] DONOHO, D. L. and GRIMES, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA* **100** 5591–5596 (electronic). [MR1981019](#)
- [10] DONOHO, D. L. and GRIMES, C. (2005). Image manifolds which are isometric to Euclidean space. *J. Math. Imaging Vision* **23** 5–24. [MR2208906](#)
- [11] GASSER, T., MÜLLER, H.-G., KÖHLER, W., MOLINARI, L. and PRADER, A. (1984). Nonparametric regression analysis of growth curves. *Ann. Statist.* **12** 210–229. [MR0733509](#)
- [12] GERVINI, D. and GASSER, T. (2004). Self-modelling warping functions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 959–971. [MR2102475](#)
- [13] GERVINI, D. and GASSER, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92** 801–820. [MR2234187](#)
- [14] GRENNANDER, U. (1950). Stochastic processes and statistical inference. *Ark. Mat.* **1** 195–277. [MR0039202](#)
- [15] HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. [MR2332269](#)
- [16] HELGASON, S. (2001). *Differential Geometry, Lie Groups, and Symmetric Spaces. Graduate Studies in Mathematics* **34**. Amer. Math. Soc., Providence, RI. [MR1834454](#)
- [17] HUCKEMANN, S. (2011). Inference on 3d Procrustes means: Tree bole growth, rank deficient diffusion tensors and perturbation models. *Scand. J. Stat.* **38** 1467–9469.
- [18] IZEM, R. and MARRON, J. S. (2007). Analysis of nonlinear modes of variation for functional data. *Electron. J. Stat.* **1** 641–676. [MR2369029](#)
- [19] JONES, M. C. and RICE, J. A. (1992). Displaying the important features of large collections of similar curves. *Amer. Statist.* **46** 140–145.
- [20] KATO, T. (1966). *Perturbation Theory for Linear Operators*. Springer, New York. [MR0203473](#)
- [21] KENDALL, D. G., BARDEN, D., CARNE, T. K. and LE, H. (1999). *Shape and Shape Theory*. Wiley, Chichester. [MR1891212](#)
- [22] KNEIP, A. and GASSER, T. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20** 1266–1305. [MR1186250](#)
- [23] KNEIP, A. and UTIKAL, K. J. (2001). Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.* **96** 519–531.
- [24] LENG, X. and MÜLLER, H. G. (2006). Time ordering of gene co-expression. *Bio-statistics* **7** 569–584.
- [25] MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. [MR0560319](#)
- [26] MÜLLER, H.-G. and YAO, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103** 1534–1544. [MR2504202](#)
- [27] RIESZ, F. and SZ-NAGY, B. (1990). *Functional Analysis*. Dover, New York. [MR1068530](#)
- [28] ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

- [29] SPELLMAN, P. T., SHERLOCK, G. and ZHANG, M. Q. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9** 3273–3297.
- [30] TANG, R. and MÜLLER, H.-G. (2009). Time-synchronized clustering of gene expression trajectories. *Biostatistics* **10** 32–45.
- [31] TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.
- [32] TUDDENHAM, R. and SNYDER, M. (1954). Physical growth of California boys and girls from birth to age 18. *California Publications on Child Development* **1** 183–364.
- [33] WANG, K. and GASSER, T. (1999). Synchronizing sample curves nonparametrically. *Ann. Statist.* **27** 439–460. [MR1714722](#)
- [34] YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. [MR2160561](#)

DEPARTMENT OF STATISTICS  
UNIVERSITY OF CALIFORNIA  
DAVIS, CALIFORNIA 95616  
USA  
E-MAIL: [dchen@wald.ucdavis.edu](mailto:dchen@wald.ucdavis.edu)  
[mueller@wald.ucdavis.edu](mailto:mueller@wald.ucdavis.edu)