

Nonlinear Manifold Representations for Functional Data

Dong Chen¹ and Hans-Georg Müller^{2 3}

Department of Statistics

University of California, Davis

One Shields Ave., Davis, CA 95616, USA

Abstract

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.

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

¹e-mail: dchen@wald.ucdavis.edu. Corresponding Author.

²e-mail: mueller@wald.ucdavis.edu.

³ Research supported in part by NSF grants DMS-0806199 and DMS-1104426.

1. INTRODUCTION

Nonlinear dimension reduction methods, such as locally linear embedding (Roweis and Saul, 2000), isometric mapping (Tenenbaum et al., 2000) and Laplacian eigenmaps (Belkin and Niyogi, 2003), 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 (Kendall et al., 1999) and the recent diffusion tensor imaging (Huckemann, 2011), 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 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 (Wang and Gasser, 1999; Gervini and Gasser, 2004), density functions (Kneip and Utikal, 2001), and functional data with pre-determined and interpretable modes (Izem and Marron, 2007). 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 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” (Kneip and Gasser, 1992). A major motivation for this proposal is that functional principal component plots, e.g., second versus first component plot, are quite often found to exhibit “horseshoe” shapes, i.e., 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 (Tenenbaum et al., 2000) 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.

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_α are open sets covering \mathcal{M} and φ_α , the coordinate maps, map the corresponding U_α onto an open subset of \mathbb{R}^d . Additional assumptions on φ_α are usually imposed in order to study the structure of \mathcal{M} (Do Carmo, 1992; Helgason, 2001).

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, i.e., the manifold can be represented by a coordinate map $\varphi : \mathbb{R}^d \rightarrow \mathcal{M} \subset L^2$, such that φ is bijective, and both φ, φ^{-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, e.g., Tenenbaum et al. (2000).

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

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

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 φ an isometric map if $L(\gamma) = L(\varphi^{-1} \circ \gamma)$ for any continuous γ , where $L(\varphi^{-1} \circ \gamma)$ is similarly defined as in (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 φ . The isometry assumption is pragmatically desirable and can be found in many approaches (Tenenbaum et al., 2000; Donoho and Grimes, 2003). Donoho and Grimes (2005) discuss conditions under which isometry holds for image data.

We use the notation $\psi \equiv \varphi^{-1}$ and refer to ψ 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 (Tenenbaum et al., 2000). More useful is the geodesic distance

$$d_g(x_1, x_2) = \inf\{L(\gamma) : \gamma(0) = x_1, \gamma(1) = x_2\}, \quad (2)$$

where the infimum is taken over all continuous paths γ 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 ψ is a representation map for \mathcal{M} . Define

$$\boldsymbol{\mu} = \mathbb{E}\{\psi(X)\}, \quad \mu^{\mathcal{M}} = \psi^{-1}(\boldsymbol{\mu}), \quad (3)$$

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 ψ is a representation map for \mathcal{M} . If ψ is isometric, the manifold mean $\mu^{\mathcal{M}}$ in (3) has the following alternative expression:*

$$\mu^{\mathcal{M}} = \arg \min_{x \in \mathcal{M}} Ed_g^2(x, X), \quad (4)$$

where d_g denotes the geodesic distance defined in (2).

The expected value in equation (4) is with respect to the probability measure that is induced by the map φ , see also Bickel and Li (2007). Equation (4) defines the Fréchet mean for geodesic distance $d_g(\cdot, \cdot)$, and therefore does not depend on the choice of the isometric map ψ . 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 $\mu(t) = EX(t)$ and $G(t, s) = \text{Cov}(X(t), X(s))$, according to Mercer's theorem (Ash and Gardner, 1975), 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 non-increasingly) and associated eigenfunctions $\{\phi_k : k \geq 1\}$,

$$G(t, s) = \sum_{k=1}^{\infty} \lambda_k \phi_k(t) \phi_k(s) \quad t, s \in \mathcal{T}. \quad (5)$$

By the Hilbert-Schmidt theorem (Grenander, 1950; Riesz and Sz-Nagy, 1990), X can be expressed in terms of the so-called Karhunen-Loève representation,

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t) \quad t \in \mathcal{T}, \quad \xi_k = \int_{\mathcal{T}} (X(t) - \mu(t)) \phi_k(t) dt, \quad (6)$$

where the ξ_k are uncorrelated random variables with mean 0 and variance λ_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 (Castro et al., 1986; Jones and Rice, 1992) are

$$X_{j,\alpha} = \mu + \alpha \lambda_j^{1/2} \phi_j, \quad j = 1, 2, \dots, \quad \alpha \in \mathbb{R}, \quad (7)$$

where factors $\lambda_j^{1/2}$ standardize the scale for different j and the functional variation in the direction of eigenfunction ϕ_j is visualized by the changing of functional shapes as α varies. However, when the functional data lie on a manifold, neither μ 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 $\psi(X) \in \mathbb{R}^d$, i.e.,

$$\text{Cov}(\psi(X)) = \sum_{j=1}^d \lambda_j^{\mathcal{M}} (\mathbf{e}_j)(\mathbf{e}_j)^T, \quad (8)$$

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

$$X_{j,\alpha}^{\mathcal{M}} = \psi^{-1}(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j), \quad j = 1, \dots, d, \quad \alpha \in \mathbb{R}, \quad (9)$$

where $\boldsymbol{\mu}$ is the mean in the d -dimensional representation space according to measure Q , as given in (3). A distinct advantage of manifold-based modes of functional variation over the principal component based version (7) is that in (9) only finitely many modes are needed, while (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 ψ 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 (9) based on representation map ψ , 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 ψ , X can be uniquely represented (due to the bijectivity of ψ) by a vector $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_d) \in \mathbb{R}^d$ in terms of

$$X = \psi^{-1}\left(\boldsymbol{\mu} + \sum_{j=1}^d \vartheta_j \mathbf{e}_j\right), \quad \vartheta_j = \langle \psi(X) - \boldsymbol{\mu}, \mathbf{e}_j \rangle, \quad j = 1, \dots, d, \quad (10)$$

where $\boldsymbol{\mu}$ and \mathbf{e}_j are defined in (3) and (8) respectively, $\langle \cdot, \cdot \rangle$ is the inner product in \mathbb{R}^d and ϑ_j are uncorrelated r.v.s with mean 0 and variance $\lambda_j^{\mathcal{M}}$. We call ϑ_j the functional manifold components (FMCs) in the representation space.

3. ESTIMATING FUNCTIONAL MANIFOLDS

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

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

Here the t_{ij} are the time points where the functions are sampled, and the $\epsilon_{ij} \in \mathbb{R}$ are i.i.d. errors with mean 0 and variance σ^2 . A first task is to find an approximation $\hat{\psi}$ to the representation map ψ 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 ψ (Izem and Marron, 2007), or one may have direct observations of $\psi(X_i)$. But in general, the representation map ψ is unknown and needs to be determined from the data.

3.1 Inferring d -dimensional manifold representations

Following Tenenbaum et al. (2000), we use the pairwise distances between observed data to obtain a map ψ that preserves the geodesic distances. Alternative approaches include LLE by Roweis and Saul (2000) and Laplacian eigenmaps by Belkin and Niyogi (2003). 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 Tenenbaum et al. (2000) 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

$$\hat{\psi} = \arg \min_{(\psi(X_1), \dots, \psi(X_n))} \sum_{i,j=1}^n \{ \|\psi(X_i) - \psi(X_j)\| - d_g(X_i, X_j) \}^2. \quad (11)$$

Here, $d_g(\cdot, \cdot)$ is the geodesic distance (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 ψ on the functions X_i , i.e., the goal is to find n vectors $\hat{\psi}(X_i) \in \mathbb{R}^d$, $i = 1, \dots, n$ that minimize (11). 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 $d_g(X_i, X_j)$ (Cox and Cox, 2001). Our asymptotic results pertain to a second step, where the assumed smoothness of ψ 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 (11), we assume that the minimization in (11) provides values on or defines the target manifold at the sample points, i.e., 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 (6) can be used to obtain fitted curves,

$$\hat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k \leq K} \hat{\xi}_{ik} \hat{\phi}_k(t). \quad (12)$$

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 ξ_{ik} are approximated by discretizing integrals

$$\hat{\xi}_{ik} = \sum_{j=2}^{n_i} \{Y_{ij} - \hat{\mu}(t_{ij})\} \hat{\phi}_k(t_{ij}) (t_{ij} - t_{i,j-1}), \quad (13)$$

or alternatively by conditional expectation (for details on these steps, see Yao et al., 2005),

$$\hat{\xi}_{ik} = \hat{\lambda}_k \hat{\phi}_{ik}^T \hat{\Sigma}_{\mathbf{Y}_i}^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i), \quad (14)$$

where $\hat{\phi}_{ik} = (\hat{\phi}_k(t_{i1}), \dots, \hat{\phi}_k(t_{in_i}))$, $(\hat{\Sigma}_{\mathbf{Y}_i})_{jl} = \hat{G}(t_{ij}, t_{il}) + \hat{\sigma}^2 \mathbf{I}(j = l)$, $1 \leq j, l \leq n_i$, $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}(t_{i1}), \dots, \hat{\mu}(t_{in_i}))$, and $\hat{\sigma}^2$ is estimated from the difference between empirical variances of Y_{ij} and $\hat{G}(t, s)$. The conditioning method (14) 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 (12), one may choose K by requiring a large fraction of variance explained (FVE), i.e.,

$$K = \min_k \left\{ k : \frac{\sum_{l \leq k} \hat{\lambda}_l}{\sum_{l=1}^{\infty} \hat{\lambda}_l} \geq 1 - \alpha \right\}, \quad (15)$$

for $\alpha = 0.05$, where the $\hat{\lambda}_l$ are estimates of the eigenvalues λ_l in (5). The resulting L^2 distances are $\|\hat{X}_i^K - \hat{X}_j^K\|_{L^2} = \left\{ \sum_{k=1}^K (\hat{\xi}_{ik} - \hat{\xi}_{jk})^2 \right\}^{\frac{1}{2}}$.

Note that alternatively to representation (12), one can also directly apply local constant or local linear smoothing to obtain smooth trajectories in the case of dense and balanced designs, e.g.,

$$\tilde{X}_i(t) = \frac{\sum_{j=1}^{n_i} \kappa_1(h_1^{-1}(t_{ij} - t)) Y_{ij}}{\sum_{j=1}^{n_i} \kappa_1(h_1^{-1}(t_{ij} - t))}, \quad (16)$$

where κ_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 can 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 (12) and (16). The estimated random trajectories, obtained through (12) or (16), 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 (Tenenbaum et al., 2000). 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 by 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

$$S(W_s, W_e) = \arg \min_{W_2, \dots, W_{m-1}} \left\{ \sum_{i=1}^{m-1} \|W_i - W_{i+1}\|_{L^2} (1 + P_\delta(W_i, W_{i+1})) : \|W_i - W_{i+1}\|_{L^2} < \epsilon \right\}. \quad (17)$$

Here the parameter ϵ limits the step length, and the penalty function P_δ 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 $\rho_{i,i+1} = \min \{ \#\{W_j : \|W_j - W_i\|_{L^2} < \epsilon\}, \#\{W_j : \|W_j - W_{i+1}\|_{L^2} < \epsilon\} \}$ and $\#$ denotes the cardinality of a set. By selecting the parameter δ , one can control the threshold of the local density of points, below which the penalty P_δ 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 δ may be selected data-adaptively by cross-validation. The choice of δ and also of the step size parameter ϵ is discussed in Section 3.3. If the manifold is very smooth, a large ϵ 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

$$\hat{d}_g(W_s, W_e) = \sum_{j=1}^{m-1} \|\hat{W}_j - \hat{W}_{j+1}\|_{L^2}, \quad (18)$$

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 mapping and representing sample trajectories

For any location $\theta \in \mathbb{R}^d$, we find $\hat{\psi}^{-1}(\theta)$ by local weighted averaging, i.e.,

$$\hat{\psi}^{-1}(\theta) = \frac{\sum_i \kappa(H^{-1}(\hat{\psi}(X_i) - \theta)) \hat{X}_i}{\sum_i \kappa(H^{-1}(\hat{\psi}(X_i) - \theta))}, \quad (19)$$

where κ is a d -dimensional kernel, such as the Epanechnikov kernel

$$\kappa(u_1, \dots, u_d) = \left(\frac{3}{4}\right)^d \prod_{k=1}^d \{(1 - u_k^2)\mathbf{I}(|u_k| < 1)\}, \quad \text{with } H = h\mathbf{I}_{d \times d}$$

for a suitably chosen bandwidth h , \hat{X}_i could be either \tilde{X}_i as in (16) or \hat{X}_i^K as in (12), and $\hat{\psi}(X_i)$ is defined after (18). We use cross-validation to select h (see section 3.3). The asymptotic properties of (19) will be discussed in section 4.

Specifically, as predictor of X_i , we propose

$$\hat{X}_i^{\mathcal{M}} = \frac{\sum_{j \neq i} \kappa(H^{-1}(\hat{\psi}(X_i) - \hat{\psi}(X_j))) \hat{X}_j}{\sum_{j \neq i} \kappa(H^{-1}(\hat{\psi}(X_i) - \hat{\psi}(X_j)))}, \quad (20)$$

borrowing strength from local neighbors in the d -dimensional representation space. This can be seen as an alternative to representation (12), where we use the FPCs and borrow strength from the whole data set to estimate functional mean and eigenbasis. As before, we note that (20) is not necessarily in \mathcal{M} , but will be in a small neighborhood asymptotically and in comparison with (12), (20) 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. The asymptotic properties are discussed in Section 4.

Definition (3) suggests to estimate the manifold mean by

$$\hat{\mu}^{\mathcal{M}} = \frac{\sum_i \kappa(H^{-1}(\hat{\psi}(X_i) - \hat{\mu})) \hat{X}_i}{\sum_i \kappa(H^{-1}(\hat{\psi}(X_i) - \hat{\mu}))}, \quad (21)$$

where $\hat{\mu} = \frac{1}{n} \sum_i \hat{\psi}(X_i)$. The FMC vectors \mathbf{e}_j defined in (8) are estimated by eigendecomposition of the sample covariance matrix of $\hat{\psi}(X_i)$, i.e., $\hat{\lambda}_j^{\mathcal{M}}$ and $\hat{\mathbf{e}}_j$ are such that

$$\sum_{j=1}^d \hat{\lambda}_j^{\mathcal{M}} (\hat{\mathbf{e}}_j)(\hat{\mathbf{e}}_j)^T = \frac{1}{n-1} \left\{ \sum_{i=1}^n \hat{\psi}(X_i) \hat{\psi}^T(X_i) - \frac{1}{n} \left(\sum_{j=1}^n \hat{\psi}(X_j) \right) \left(\sum_{j=1}^n \hat{\psi}(X_j) \right)^T \right\}, \quad (22)$$

where the $\hat{\lambda}_j^{\mathcal{M}}$ are ordered to be non-increasing in j . From (9) and (19), we obtain estimates of the manifold modes as

$$\hat{X}_{j,\alpha}^{\mathcal{M}} = \frac{\sum_i \kappa(H^{-1}\{\hat{\psi}(X_i) - \hat{\mu} - \alpha(\hat{\lambda}_j^{\mathcal{M}})^{1/2} \hat{\mathbf{e}}_j\}) \hat{X}_i}{\sum_i \kappa(H^{-1}\{\hat{\psi}(X_i) - \hat{\mu} - \alpha(\hat{\lambda}_j^{\mathcal{M}})^{1/2} \hat{\mathbf{e}}_j\})}, \quad j = 1, \dots, d, \quad \alpha \in \mathbb{R}. \quad (23)$$

3.3 Selection of auxiliary parameters

We use 10-fold cross-validation to simultaneously choose the step size ϵ , the truncation parameter δ , and the smoothing bandwidth h (see sections 3.1 and 3.2). The number of candidates for ϵ and δ is kept small so that the cross-validation procedure runs reasonably fast. Candidates for the step size ϵ are the median distance of the 5th, the 8th and the 12th nearest neighbor; those for δ 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 ϵ, δ to determine $\hat{\psi}(X_i)$, followed by estimation of $X_i = \psi^{-1}(\vartheta_i)$ for X_i in the validation set, using (19) 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 $Y_{il} = X_i(t_{il}) + \epsilon_{il}$ is the observed value of trajectory X_i at time t_{ij} . The cross-validation choice is the minimizer of $\text{MSPE}(\epsilon, h, \delta) = \frac{\sum_{k=1}^{10} \text{SSPE}_k}{\sum_{i=1}^n n_i}$.

Following Tenenbaum et al. (2000), the intrinsic dimension d can be chosen by the $1 - \beta$ fraction of distances explained (FDE), that is,

$$d = \min_p \left\{ p : \frac{\|\hat{D}^p - D\|_F}{\|D\|_F} < \beta \right\}, \quad (24)$$

where $\beta = 0.05$ and D, \hat{D}^p are n by n distance matrixes with $D_{ij} = \hat{d}_g(X_i, X_j)$ as in (18), $\hat{D}_{ij}^p = \|\hat{\psi}^p(X_i) - \hat{\psi}^p(X_j)\|$, where $\hat{\psi}^p$ denotes the MDS solution (11) in \mathbb{R}^p , and $\|\cdot\|_F$ is the matrix Frobenius norm, $\|D\|_F = \{\sum_{i,j} D_{ij}^2\}^{\frac{1}{2}}$. Note that $\|\hat{D}^p - D\|_F$ is the square root of the minimized value of (11).

4. ASYMPTOTIC PROPERTIES

We provide the specific convergence rate of \hat{X}_i^K , defined in (12), 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 (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 let $r_n = \max\{\frac{1}{\sqrt{nh_G^2}}, \frac{1}{\sqrt{nh_\mu}}, \frac{1}{\sqrt{nh_V}}\}$. If there are infinitely many nonzero eigenvalues λ_k in (5), which are all of multiplicity one, then for sequences $K = K_n \rightarrow \infty$, subject to $r_n K^{\alpha_2 + \frac{1}{2}} \rightarrow 0$, where α_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 (5) and \hat{G} is defined after (12), it holds that*

$$\|\hat{X}_i^K - X_i\|_{L^2} = O_p(r_n K^{\alpha_2 + \frac{1}{2}} + K^{-\frac{1}{2}(\alpha_1 - 1)}), \quad (25)$$

for \hat{X}_i^K defined in (12), where α_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 (25) 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 λ_k in (5), it can be shown that the rate in (25) simply becomes $O_p(r_n)$. Next we discuss the convergence of the estimates that appear in (22).

Proposition 4. *Under (B1)-(B2) in the Appendix,*

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_p(v_n + \frac{1}{\sqrt{n}}), \quad (26)$$

where $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ are defined in (3) and (21), 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

$$\|\hat{\mathbf{e}}_j - \mathbf{e}_j\| = O_p\left(v_n + \frac{1}{\sqrt{n}}\right), \quad (27)$$

$$|\hat{\lambda}_j^{\mathcal{M}} - \lambda_j^{\mathcal{M}}| = O_p\left(v_n + \frac{1}{\sqrt{n}}\right), \quad (28)$$

where $\lambda_j^{\mathcal{M}}$, \mathbf{e}_j , $\hat{\mathbf{e}}_j$ and $\hat{\lambda}_j^{\mathcal{M}}$ are defined in (8) and (22), 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 $h^{-(d+1)}Ev_n \rightarrow 0$. Then $\hat{\psi}^{-1}(\boldsymbol{\theta})$ defined in (19), using $\hat{X}_i = \hat{X}_i^K$, is a consistent estimate of $\psi^{-1}(\boldsymbol{\theta})$. Specifically, defining $T_\phi^K = \{\sum_{k>K} \xi_k^2\}^{\frac{1}{2}}$ where $\xi_k = \int (X - EX)\phi_k$ and the orthonormal basis $\{\phi_k : k \geq 1\}$ is given in (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

$$\|\hat{\psi}^{-1}(\boldsymbol{\theta}) - \psi^{-1}(\boldsymbol{\theta})\|_{L^2} = O_p\left(h^2 + \frac{1}{\sqrt{nh^d}} + \frac{v_n}{h} + R_K(\boldsymbol{\theta}) + K^{\alpha_2 + \frac{1}{2}}r_n\right), \quad (29)$$

where r_n , α_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 $K^{\alpha_2 + \frac{1}{2}}r_n$ 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 ψ are known, reflecting an intrinsically d -dimensional smoothing problem. Related findings are discussed in Bickel and Li (2007).

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 $f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) > 0$ and that h is chosen as in Theorem 1. Then the estimated manifold modes $\hat{X}_{j,\alpha}^{\mathcal{M}}$ as in (23), substituting $\hat{X}_i = \hat{X}_i^K$, are consistent. Specifically,

$$\|\hat{X}_{j,\alpha}^{\mathcal{M}} - X_{j,\alpha}^{\mathcal{M}}\|_{L^2} = O_p\left(h^2 + \frac{1}{\sqrt{nh^d}} + \frac{v_n}{h} + \frac{1}{\sqrt{nh}} + R_K + K^{\alpha_2 + \frac{1}{2}}r_n\right), \quad (30)$$

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

An immediate consequence of these results is that the manifold representation given in (10) provides a consistent representation of all random functions in the functional manifold.

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

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

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 μ , which has two peaks. The time warping function h_α is generated from the cumulative Beta distribution family and α is a random parameter, $\alpha = \max(-1, Z)$, where $Z \sim N(0, 0.09)$.

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

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

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

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

a mixture of two peaks with randomly varying centers, where $\alpha \sim N(0, 1)$ and $\beta \sim N(0, 1)$. 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 (1), the condition $\int_{\theta_k^0}^{\theta_k^1} \|\frac{\partial X}{\partial \theta_k}(t)\|_{L^2} d\theta_k \equiv \theta_k^1 - \theta_k^0$ 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 ψ 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 ψ 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 .

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 α 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 (24). 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 non-isomorphic, 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 (Cox and Cox, 2001) 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 non-isometric manifold \mathcal{M}_3 are not sufficient. Thus, the non-isometric 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 non-isometric 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 non-isometric 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 non-isometric 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 non-isometric 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 $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$ 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 σ^2 is such that the signal-to-noise ratio R is 0.1 or 0.5. We estimated manifold means $\mu^{\mathcal{M}}$ (3), manifold modes of functional variation $X_{j,\alpha}^{\mathcal{M}}$ (9) and obtained predictions $\hat{X}_i^{\mathcal{M}}$ (20), which were compared with predictions obtained by functional principal component analysis.

Results for a simulation run are shown in Figures 2–4 for manifolds $\mathcal{M}_1 - \mathcal{M}_3$, respectively. The estimated manifold means are seen to be close to the corresponding intrinsic means, i.e., 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.

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 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 (12), resulting in \hat{X}_i^L , as well as the proposed new estimates $\hat{X}_i^{\mathcal{M}}$ (20). For \hat{X}_i^L , we estimate the FPCs (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 one, 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 L and d up to 5.

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, i.e., 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 ϵ in P-ISOMAP, defined in eq.

(17), on mean squared prediction errors is demonstrated in Table 4. Here d is fixed as the intrinsic dimension (1 for \mathcal{M}_1 and 2 for $\mathcal{M}_2, \mathcal{M}_3$), while δ and h are optimized by cross-validation for each ϵ . We then select ϵ 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 ϵ , as long as it is in medium range. A good overall choice is median distance of 8th nearest neighbors. When ϵ 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. In practice, we therefore impose a lower bound on ϵ to ensure that the number of data that are not connected to other points when connecting through ϵ -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 representing important growth features adequately (Kneip and Gasser, 1992; Gervini and Gasser, 2005). Since phase variation introduces non-linear 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 (Tuddenham and Snyder, 1954). 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 (Gasser et al., 1984), 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 $\hat{\mu}^{\mathcal{M}}$ (21). 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), defined in (24), for different dimensions d are shown in Table 5. The MSPE of \hat{X}_j^L is minimized at $L = 5$, with $L = 2$ already a quite good choice.

We find that $\hat{X}_j^{\mathcal{M}}$ 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 (20) with $K = 4$, applying criterion (15). 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 $\hat{X}_i^{\mathcal{M}}$ using $L = d = 2$. The two bottom panels of Figure 5 illustrate the comparison of estimated manifold 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 Spellman et al. (1998). There are 6178 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 co-expressed 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 (Leng and Müller, 2006; Tang and Müller, 2009).

The data we study consist of 90 genes that have been identified by biological methods (Spellman et al., 1998). 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 (red) and S (cyan) groups, and one peak for G2/M (green) and M/G1 (blue) groups, while the trajectories for the S/G2 (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 $\hat{\mu}^M$ (21) (middle left panel of Figure 6) is seen to fall within the G1 group (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 (12) and \hat{X}_i^M (20), using preliminary estimators \hat{X}_i^K with $K = 4$ in (20). 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 (black), is well represented by the variation across this manifold mode.

6.3 Human mortality across countries

The death rates derived from current lifetable 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 pre-smoothed these data, following (16). The resulting MSPEs for $\hat{X}_i^{\mathcal{M}}$ (20) and \hat{X}_i^L (12) 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 $\hat{\mu}^{\mathcal{M}}$ (21) 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 (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.

6. 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 (De Silva and Tenenbaum, 2003), 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 identi-

fied 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 non-linear, 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 Tenenbaum et al. (2000).

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

A.1 Assumptions

(A1) The bandwidths h_μ , h_v , h_G for estimating $\mu(t)$, σ^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 κ_μ for the mean function μ and κ_G for the covariance function G in section 3.1 are absolutely integrable, i.e., $\int |\kappa_\mu(t)|dt < \infty$ and $\int \int |\kappa_G(t, s)|dtds < \infty$.

- (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 ψ converge uniformly on the sample space, i.e., $E v_n \rightarrow 0$ 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 κ satisfies $\int \kappa(\mathbf{u}) d\mathbf{u} = 1$, $\det(\int \kappa(\mathbf{u}) \mathbf{u} \mathbf{u}^T d\mathbf{u}) < \infty$, $\int \kappa^2(\mathbf{u}) d\mathbf{u} < \infty$, $\kappa(\mathbf{u}) = \kappa(-\mathbf{u})$, 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, i.e. 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} .

A.2 Proofs

Proof of Proposition 1. Note that $\|\psi(x) - \psi(y)\| \equiv d_g(x, y)$ for any $x, y \in \mathcal{M}$ if ψ is isometric, because $\|\psi(x) - \psi(y)\|$ is the shortest distance between $\psi(x)$ and $\psi(y)$ in \mathbb{R}^d . Then $\text{Ed}_g^2(x, X) \equiv \text{E}\|\psi(x) - \psi(X)\|^2$ for any $x \in \mathcal{M}$. Since $\text{E}\{\psi(X)\} = \arg \min_{\theta \in \mathbb{R}^d} \text{E}\|\theta - \psi(X)\|^2$, bijectivity of ψ implies $\arg \min_{x \in \mathcal{M}} \text{Ed}_g^2(x, X) = \psi^{-1}(\text{E}\{\psi(X)\})$.

Proof of Proposition 2. Define $\pi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ as $\pi = \tilde{\psi} \circ \psi^{-1}$. It suffices to show that if π is isometric, π must be a rigid transformation, i.e., $\pi(\theta) = \theta_0 + A\theta$ for any $\theta \in \mathbb{R}^d$, a constant vector $\theta_0 \in \mathbb{R}^d$ and an orthogonal matrix A . This can be done in the following two steps: Let e_1, \dots, e_d be an orthogonal basis of \mathbb{R}^d , then $\tilde{e}_k = \pi(e_k) - \pi(0)$, $k = 1, \dots, d$, must also be an orthogonal basis of \mathbb{R}^d since $\|\tilde{e}_k\| \equiv 1$ and $\|\tilde{e}_k - \tilde{e}_{k'}\| \equiv \sqrt{2}$ for any $k \neq k'$; for any vector $\theta = \sum_{k=1}^d \alpha_k e_k \in \mathbb{R}^d$, it can be shown that $\langle \pi(\theta) - \pi(0), \tilde{e}_k \rangle = \alpha_k$ for $1 \leq k \leq d$ by solving the equations $\|\pi(\theta) - \pi(e_k)\| = \|\theta - e_k\|$, $1 \leq k \leq d$.

Proof of Proposition 3. Note that for \hat{X}_i^K defined in (12), it holds that

$$\|\hat{X}_i^K - X_i\|_{L^2} \leq \|\hat{X}_i^K - \sum_{k \leq K} \xi_{ik} \phi_k\|_{L^2} + \|\sum_{k > K} \xi_{ik} \phi_k\|_{L^2}. \quad (31)$$

Since the ϕ_k are orthonormal and the ξ_{ik} have mean 0, for the second term on the r.h.s., $\text{E}\|\sum_{k > K} \xi_{ik} \phi_k\|_{L^2}^2 = \sum_{k > K} \lambda_k$. Under condition (A4), $\sum_{k > K} \lambda_k < C_1 K^{-(\alpha_1 - 1)}$. Using Chebyshev's inequality,

$$\|\sum_{k > K} \xi_{ik} \phi_k\|_{L^2} = O_p(K^{-\frac{1}{2}(\alpha_1 - 1)}). \quad (32)$$

For the first term on the r.h.s. of (31),

$$\|\hat{X}_i^K - X_i^K\|_{L^2} \leq \|\hat{\mu} - \mu\|_{L^2} + \|\sum_{k \leq K} (\hat{\xi}_{ik} - \xi_{ik}) \hat{\phi}_k\|_{L^2} + \|\sum_{k \leq K} \xi_{ik} (\hat{\phi}_k - \phi_k)\|_{L^2}. \quad (33)$$

Next we prove that:

$$\|\hat{\mu} - \mu\|_{L^2} = O_p(r_n), \quad (34)$$

$$\|\sum_{k \leq K} (\hat{\xi}_{ik} - \xi_{ik}) \hat{\phi}_k\|_{L^2} = O_p(K^{\alpha_2 + \frac{1}{2}} r_n), \quad (35)$$

$$\left\| \sum_{k \leq K} \xi_{ik}(\hat{\phi}_k - \phi_k) \right\|_{L^2} = O_p(K^{\alpha_2 + \frac{1}{2}} r_n). \quad (36)$$

Here (34) follows from

$$\|\hat{\mu} - \mu\|_\infty = O_p(r_n) \quad (37)$$

in Yao et al. (2005). To show (35), we introduce the following abbreviations:

$$\begin{aligned} Q_i^{(1)} &= \sum_{j=2}^{n_i} \{X_i(t_{ij}) - \mu(t_{ij})\}^2 \tau_{ij}; & Q_{ik}^{(2)} &= \sum_{j=2}^{n_i} \phi_k^2(t_{ij}) \tau_{ij}; & Q_{ik}^{(3)} &= \left\{ \sum_{j=2}^{n_i} \epsilon_{ij} \phi_k(t_{ij}) \tau_{ij} \right\}^2; \\ Q_i^{(4)} &= \sum_{j=2}^{n_i} \epsilon_{ij}^2 \tau_{ij}; & Z_k^{(1)} &= \|\hat{\phi}_k - \phi_k\|_\infty^2; & Z^{(2)} &= \|\hat{\mu} - \mu\|_\infty^2; & Z_{ik}^{(3)} &= \|\{(X_i - \mu)\phi_k\}'\|_\infty^2; \end{aligned} \quad (38)$$

and argue that

$$\begin{aligned} Q_i^{(1)} &= O_p(1); & \sum_{k \leq K} Q_{ik}^{(2)} &= O_p(K); & \sum_{k \leq K} Q_{ik}^{(3)} &= O_p(K r_n^2); & Q_i^{(4)} &= O_p(1); \\ \sum_{k \leq K} Z_k^{(1)} &= O_p(K^{2\alpha_2 + 1} r_n^2); & Z^{(2)} &= O_p(r_n^2); & \sum_{k \leq K} Z_{ik}^{(3)} &= O_p(K^{\alpha_2 + 1}). \end{aligned} \quad (39)$$

Here we use the fact that if a function f is differentiable, then

$$\left| \int_a^b f(t) dt - \sum_{j=1}^N f(t_j)(t_j - t_{j-1}) \right| \leq 2 \|f'\|_\infty \max_j |t_j - t_{j-1}| (b - a), \quad (40)$$

where t_j , $j = 1, \dots, N$, is any partition of $[a, b]$. Without loss of generality, we assume that the interval considered is $[0, 1]$. Let $f = (X_i - \mu)^2$, then $\|f'\|_\infty \leq 2 \|X_i - \mu\|_\infty \cdot \|X_i' - \mu'\|_\infty = O_p(1)$ under (A5). So

$$Q_i^{(1)} \leq \|X_i - \mu\|_{L^2}^2 + 4\tau^* \|X_i - \mu\|_\infty \cdot \|X_i' - \mu'\|_\infty = O_p(1), \quad (41)$$

where τ^* is defined in (A3). Again, if $f = \phi_k^2$, then $\|f'\|_\infty = \|2\phi_k \phi_k'\|_\infty \leq \frac{2\sqrt{C_3 C_4}}{\lambda_k}$ because $\lambda_k \phi_k^2(t) \leq C_3$ and $\lambda_k \phi_k'^2(t) \leq C_4$ uniformly under condition (A5). Then

$$Q_{ik}^{(2)} \leq 1 + \frac{4\sqrt{C_3 C_4} \tau^*}{\lambda_k}. \quad (42)$$

Since $\tau^* = O_p(r_n^2)$ and $\frac{1}{\lambda_k} \leq \frac{1}{\delta_k} \leq C_2 k^{\alpha_2}$, where $\delta_k = \min_{j \leq k} (\lambda_j - \lambda_{j+1})$ under (A3) and (A4), we have $\sum_{k \leq K} \frac{\tau^*}{\lambda_k} = O_p(K^{\alpha_2 + 1} r_n^2)$, which can be omitted because K is selected such that $K^{2\alpha_2 + 1} r_n^2 \rightarrow 0$ in Lemma 1. For $Q_{ik}^{(3)}$, since

$$EQ_{ik}^{(3)} = \sigma^2 \sum_{j=2}^{n_i} \phi_k^2(t_{ij}) \tau_{ij}^2 \leq \sigma^2 \tau^* \sum_{j=2}^{n_i} \phi_k^2(t_{ij}) \tau_{ij} \leq \sigma^2 \tau^* \left(1 + \frac{4\sqrt{C_3 C_4} \tau^*}{\lambda_k}\right), \quad (43)$$

analogously to $Q_{ik}^{(2)}$, we infer $\sum_{k \leq K} Q_{ik}^{(3)} = O_p(Kr_n^2)$. The order of $Q_i^{(4)}$ is obvious by calculating $EQ_i^{(4)}$. To check $\sum_{k \leq K} Z_k^{(1)}$, we use a uniform convergence result (Hall and Horowitz, 2007; Müller and Yao, 2008):

$$\|\hat{\phi}_k - \phi_k\|_\infty = O_p\left(\frac{r_n}{\delta_k}\right), \quad k = 1, \dots, K_0,$$

where $\delta_k = \min_{j \leq k} (\lambda_j - \lambda_{j+1})$ and K_0 is defined in Lemma 1. Under assumption (A4), we have $\delta_k > C_2 k^{-\alpha_2}$ and $\sum_{k \leq K} \frac{1}{\delta_k^2} = O(K^{2\alpha_2+1})$. So provided that $K \leq K_0$, one finds $\sum_{k \leq K} Z_k^{(1)} = O_p(K^{2\alpha_2+1} r_n^2)$. The order of $Z^{(2)}$ is a consequence of (37). For $Z_{ik}^{(3)}$, observe

$$Z_{ik}^{(3)} \leq 2\|X_i - \mu\|_\infty^2 \cdot \|\phi'_k\|_\infty^2 + 2\|X'_i - \mu'\|_\infty^2 \cdot \|\phi_k\|_\infty^2. \quad (44)$$

Then $\sum_{k \leq K} Z_{ik}^{(3)} = O_p(K^{\alpha_2+1})$ analogously to $Q_{ik}^{(2)}$.

For (35), due to the orthonormality of $\hat{\phi}_k$,

$$\left\| \sum_{k \leq K} (\hat{\xi}_{ik} - \xi_{ik}) \hat{\phi}_k \right\|_{L^2}^2 = \sum_{k \leq K} (\hat{\xi}_{ik} - \xi_{ik})^2.$$

Letting

$$\begin{aligned} \hat{\eta}_{ik} &= \sum_{j=2}^{n_i} \{X_i(t_{ij}) - \hat{\mu}(t_{ij})\} \hat{\phi}_k(t_{ij})(t_{ij} - t_{i,j-1}); & \tilde{\eta}_{ik} &= \sum_{j=2}^{n_i} \{X_i(t_{ij}) - \mu(t_{ij})\} \phi_k(t_{ij})(t_{ij} - t_{i,j-1}); \\ \hat{\varepsilon}_{ik} &= \sum_{j=2}^{n_i} \{\epsilon_{ij} \hat{\phi}_k(t_{ij})(t_{ij} - t_{i,j-1})\}, \end{aligned}$$

$$(\hat{\xi}_{ik} - \xi_{ik})^2 \leq 4(\hat{\eta}_{ik} - \tilde{\eta}_{ik})^2 + 4(\tilde{\eta}_{ik} - \xi_{ik})^2 + 2(\hat{\varepsilon}_{ik})^2. \quad (45)$$

For the first term on the r.h.s.,

$$(\hat{\eta}_{ik} - \tilde{\eta}_{ik})^2 \leq 4Q_i^{(1)} Z_k^{(1)} + 4Z_k^{(1)} Z^{(2)} + 2Q_{ik}^{(2)} Z^{(2)},$$

and for the second term on the r.h.s. of (45), by (40), $(\tilde{\eta}_{ik} - \xi_{ik})^2 \leq 4Z_{ik}^{(3)} \tau^{*2}$. The last term can be bounded by

$$\hat{\varepsilon}_{ik}^2 \leq 2 \left\{ \sum_{j=2}^{n_i} \epsilon_{ij} \phi_k(t_{ij}) \tau_{ij} \right\}^2 + 2 \left\{ \sum_{j=2}^{n_i} \epsilon_{ij} [\hat{\phi}_k(t_{ij}) - \phi_k(t_{ij})] \tau_{ij} \right\}^2 \leq 2Q_{ik}^{(3)} + 2Z_k^{(1)} Q_i^{(4)}.$$

Combining these bounds,

$$\begin{aligned} \left\| \sum_{k \leq K} (\hat{\xi}_{ik} - \xi_{ik}) \hat{\phi}_k \right\|_{L^2}^2 &\leq 16 Q_i^{(1)} \sum_{k \leq K} Z_k^{(1)} + 8 \sum_{k \leq K} Q_{ik}^{(2)} Z^{(2)} + 16 \sum_{k \leq K} Z_k^{(1)} Z^{(2)} \\ &\quad + 16 \sum_{k \leq K} Z_{ik}^{(3)} \tau^{*2} + 4 \sum_{k \leq K} Q_{ik}^{(3)} + 4 \sum_{k \leq K} Z_k^{(1)} Q_i^{(4)}, \end{aligned} \quad (46)$$

and (39) implies (35).

For (36), by the Cauchy-Schwarz inequality and (39),

$$\begin{aligned} \left\| \sum_{k \leq K} \xi_{ik} (\hat{\phi}_k - \phi_k) \right\|_{L^2} &\leq \left\{ \sum_{k \leq K} \xi_{ik}^2 \right\}^{\frac{1}{2}} \left\{ \sum_{k \leq K} \|(\hat{\phi}_k - \phi_k)\|_{L^2}^2 \right\}^{\frac{1}{2}} \\ &\leq \|X_i\|_{L^2} \left(\sum_{k \leq K} Z_k^{(1)} \right)^{\frac{1}{2}} = O_p(K^{\alpha_2 + \frac{1}{2}} r_n). \end{aligned} \quad (47)$$

Combining (31)-(36) implies (25).

Proof of Proposition 4. Let $\hat{\boldsymbol{\vartheta}}_i = \hat{\psi}(X_i)$ and $\boldsymbol{\vartheta}_i = \psi(X_i)$. By (B1), (B2) and the Central Limit Theorem,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \left\| \frac{1}{n} \sum_{i=1}^n (\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i) \right\| + \left\| \frac{1}{n} \sum_{i=1}^n (\boldsymbol{\vartheta}_i - \mathbf{E}\boldsymbol{\vartheta}) \right\| = O_p(v_n + \frac{1}{\sqrt{n}}).$$

To show (27), define $\hat{\Sigma}_n = \frac{1}{n-1} \sum_{i=1}^n \hat{\boldsymbol{\vartheta}}_i \hat{\boldsymbol{\vartheta}}_i^T - \frac{n}{n-1} \bar{\boldsymbol{\vartheta}} \bar{\boldsymbol{\vartheta}}^T$ and $\Sigma_n = \frac{1}{n-1} \sum_{i=1}^n \boldsymbol{\vartheta}_i \boldsymbol{\vartheta}_i^T - \frac{n}{n-1} \tilde{\boldsymbol{\vartheta}} \tilde{\boldsymbol{\vartheta}}^T$, where $\bar{\boldsymbol{\vartheta}} = \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol{\vartheta}}_i$ and $\tilde{\boldsymbol{\vartheta}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\vartheta}_i$. For two d by d matrixes A and B , define the Frobenius product $\langle A, B \rangle_F = \text{tr}(AB^T)$ and the Frobenius norm $\|A\|_F^2 = \langle A, A \rangle_F$. Then

$$\begin{aligned} \|\hat{\Sigma}_n - \Sigma\|_F &\leq \|\hat{\Sigma}_n - \Sigma_n\|_F + \|\Sigma_n - \Sigma\|_F \leq \frac{1}{n-1} \left\| \sum_{i=1}^n (\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i) (\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i)^T \right\|_F + \quad (48) \\ &\quad \frac{2}{n-1} \left\| \sum_{i=1}^n (\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i) \boldsymbol{\vartheta}_i^T \right\|_F + \frac{n}{n-1} \|(\bar{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}}) (\bar{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}})^T\|_F + \frac{2n}{n-1} \|(\bar{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}}) \tilde{\boldsymbol{\vartheta}}^T\|_F + \|\Sigma_n - \Sigma\|_F. \end{aligned}$$

The first term on the last r.h.s. is $O_p(v_n^2)$ because $\sum_{i=1}^n \|(\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i) (\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i)^T\|_F = \sum_{i=1}^n \|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\|^2$; for the second term, $\frac{1}{n} \sum_{i=1}^n \|(\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i) \boldsymbol{\vartheta}_i^T\|_F \leq \left\{ \left(\frac{1}{n} \sum_{i=1}^n \|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\|^2 \right) \left(\frac{1}{n} \sum_{i=1}^n \|\boldsymbol{\vartheta}_i\|^2 \right) \right\}^{\frac{1}{2}} = O_p(v_n)$; similarly, the third term is $O_p(v_n^2)$ and the fourth is $O_p(v_n)$; the last term is $O_p(\frac{1}{\sqrt{n}})$ (Mardia et al., 1979). Thus $\|\hat{\Sigma}_n - \Sigma\|_F = O_p(v_n + \frac{1}{\sqrt{n}})$.

Denote by $\lambda_1(A) \geq \lambda_2(A) \geq \dots \geq \lambda_d(A)$ the eigenvalues of a d by d symmetric matrix A , and by $e_1(A), \dots, e_d(A)$ the corresponding eigenvectors (supposing they are unique).

Let $\|A\|_2 = \max\{|\lambda_1(A)|, \dots, |\lambda_p(A)|\}$, and define $P_j(A)$ as the eigenprojection matrix, i.e., $P_j(A) = e_j(A)e_j^T(A)$. If $\lambda_j(\Sigma)$ is of multiplicity one, $d_j = \min\{\lambda_{j-1}(\Sigma) - \lambda_j(\Sigma), \lambda_j(\Sigma) - \lambda_{j+1}(\Sigma)\} > 0$. Then the Taylor expansion of $P_k(\hat{\Sigma}_n)$ (Kato, 1966),

$$P_j(\hat{\Sigma}_n) - P_j(\Sigma) = P_j(\Sigma)(\hat{\Sigma}_n - \Sigma)(\Sigma - \lambda_j I)^+ + (\Sigma - \lambda_j I)^+(\hat{\Sigma}_n - \Sigma)P_j(\Sigma) + E_n, \quad (49)$$

where A^+ is the Moore-Penrose generalized inverse of A and $\|E_n\|_2 \leq (2\|\hat{\Sigma}_n - \Sigma\|_2/d_j)^2(1 - 2\|\hat{\Sigma}_n - \Sigma\|_2/d_j)^{-1}$, satisfies $\|P_j(\hat{\Sigma}_n) - P_j(\Sigma)\|_F = O_p(\|\hat{\Sigma}_n - \Sigma\|_F)$. Now (27) follows from

$$\|P_j(\hat{\Sigma}_n) - P_j(\Sigma)\|_F^2 = 2\{1 - (e_j^T(\hat{\Sigma}_n)e_j(\Sigma))\}^2 \geq 2\|e_j(\hat{\Sigma}_n) - e_j(\Sigma)\|^2. \quad (50)$$

Lastly, we conclude (28) from

$$|\hat{\lambda}_j^{\mathcal{M}} - \lambda_j^{\mathcal{M}}| = |e_j^T(\hat{\Sigma}_n)\hat{\Sigma}_n e_j(\hat{\Sigma}_n) - e_j^T(\Sigma)\Sigma e_j(\Sigma)| = O_p(\|e_j(\hat{\Sigma}_n) - e_j(\Sigma)\| + \|\hat{\Sigma}_n - \Sigma\|_F). \quad (51)$$

Proof of Theorem 1 and Corollary 1. We only show (30). To simplify notations, define $\hat{w}_i^{j,\alpha} = \frac{1}{nh^d}\kappa(H^{-1}(\hat{\boldsymbol{\vartheta}}_i - \hat{\boldsymbol{\mu}} - \alpha(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}}\hat{\mathbf{e}}_j))$ and $w_i^{j,\alpha} = \frac{1}{nh^d}\kappa(H^{-1}(\boldsymbol{\vartheta}_i - \boldsymbol{\mu} - \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j))$, where $\hat{\boldsymbol{\vartheta}}_i = \hat{\psi}(X_i)$ and $\boldsymbol{\vartheta}_i = \psi(X_i)$. Then

$$\begin{aligned} \|\hat{X}_{j,\alpha}^{\mathcal{M}} - X_{j,\alpha}^{\mathcal{M}}\|_{L^2} &\leq \left\| \frac{\sum_i w_i^{j,\alpha} X_i}{\sum_i w_i^{j,\alpha}} - X_{j,\alpha}^{\mathcal{M}} \right\|_{L^2} + \left\| \sum_{i=1}^n \left(\frac{\hat{w}_i^{j,\alpha}}{\sum_l \hat{w}_l^{j,\alpha}} - \frac{w_i^{j,\alpha}}{\sum_l w_l^{j,\alpha}} \right) X_i \right\|_{L^2} \\ &\quad + \left\| \frac{\sum_i w_i^{j,\alpha} (\hat{X}_i^K - X_i)}{\sum_i \hat{w}_i^{j,\alpha}} \right\|_{L^2} + \left\| \frac{\sum_i (\hat{w}_i^{j,\alpha} - w_i^{j,\alpha}) (\hat{X}_i^K - X_i)}{\sum_i \hat{w}_i^{j,\alpha}} \right\|_{L^2}. \end{aligned} \quad (52)$$

For the first term on the r.h.s., let $g_1 = \sum_i w_i^{j,\alpha}$ and $g_2 = \sum_i w_i^{j,\alpha} X_i$. By standard arguments for kernel smoothing, under conditions (C1) and (C2), $g_1 = f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) + O_p(h^2 + \frac{1}{\sqrt{nh^d}})$. For g_2 , we use $X_i = \psi^{-1}(\boldsymbol{\vartheta}_i)$ and require ψ^{-1} to be twice Fréchet differentiable as defined in (C3), which implies that $\psi^{-1}(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j + H\mathbf{v}) = \psi^{-1}(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) + hJ_1^T \mathbf{v} + R_1$ where J_1 is a d -vector consisting of L^2 functions and R_1 is a residual term. We aim to show that $\|g_2 - f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j)X_{j,\alpha}^{\mathcal{M}}\|_{L^2} = O_p(h^2 + \frac{1}{\sqrt{nh^d}})$, which is guaranteed if $E\|g_2 - f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j)X_{j,\alpha}^{\mathcal{M}}\|_{L^2}^2 = O(h^4 + \frac{1}{nh^d})$. To prove this, consider a complete L^2 basis $\{\varphi_k, k = 1, 2, \dots\}$. Suppose $\omega_k = \langle g_2 - f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j)X_{j,\alpha}^{\mathcal{M}}, \varphi_k \rangle_{L^2}$ where $\langle \cdot, \cdot \rangle_{L^2}$ is the

inner product in L_2 , then $\mathbb{E}\|g_2 - f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j)X_{j,\alpha}^{\mathcal{M}}\|_{L^2}^2 = \sum_k (\mathbb{E}\omega_k)^2 + \sum_k \text{Var}(\omega_k)$. To bound $\sum_k (\mathbb{E}\omega_k)^2$, note that

$$\begin{aligned} \mathbb{E}\langle g_2, \varphi_k \rangle_{L^2} &= \frac{1}{h^d} \int \kappa(H^{-1}(\mathbf{u} - \boldsymbol{\mu} - \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j)) \langle \psi^{-1}(\mathbf{u}), \varphi_k \rangle_{L^2} f(\mathbf{u}) d\mathbf{u} \\ &= \int \kappa(\mathbf{v}) \langle \psi^{-1}(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) + hJ_1^T \mathbf{v} + R_1, \varphi_k \rangle_{L^2} \{f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) + hJ_2 \mathbf{v} + R_2\} d\mathbf{v} \\ &= \langle f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) \psi^{-1}(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) + R_3, \varphi_k \rangle_{L^2}, \end{aligned}$$

where J_2 is a d -vector consisting of partial derivatives of f , and $R_3 = f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) \int R_1 \kappa(\mathbf{v}) d\mathbf{v} + \psi^{-1}(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) \int R_2 \kappa(\mathbf{v}) d\mathbf{v}$. Under (C1) and (C2), $\|R_1\|_{L^2} \leq C_1 h^2 \|v\|^2$, $\|R_2\|_{L^2} \leq C_2 h^2 \|v\|^2$, and $\|R_3\|_{L^2} = O(h^2)$, implying $\sum_k (\mathbb{E}\omega_k)^2 = O(h^4)$ under (C2). For the variance part, using (C2),

$$\begin{aligned} \sum_k \text{Var}\langle g_2, \varphi_k \rangle_{L^2} &= n \sum_k \text{Var}\langle \frac{1}{nh^d} \kappa(H^{-1}(\mathbf{u} - \boldsymbol{\mu} - \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j)) \psi^{-1}(\mathbf{u}), \varphi_k \rangle_{L^2} \\ &= \frac{1}{nh^{2d}} \sum_k \int \kappa^2(H^{-1}(\mathbf{u} - \boldsymbol{\mu} - \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j)) \langle \psi^{-1}(\mathbf{u}), \varphi_k \rangle_{L^2}^2 f(\mathbf{u}) d\mathbf{u} - \frac{1}{n} \sum_k \{\mathbb{E}\langle g_2, \varphi_k \rangle_{L^2}\}^2 \\ &= \frac{1}{nh^d} \int \kappa^2(\mathbf{v}) \sum_k \left\{ \langle \psi^{-1}(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) + hJ_1^T \mathbf{v} + R_1, \varphi_k \rangle_{L^2}^2 \right. \\ &\quad \left. \langle f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j) + hJ_2 \mathbf{v} + R_2 \rangle_{L^2} d\mathbf{v} + O\left(\frac{1}{n}\right) = O\left(\frac{1}{nh^d}\right). \right. \end{aligned}$$

Thus the first term on the right side of (52) is $O_p(h^2 + \frac{1}{\sqrt{nh^d}})$.

For the second term, note that

$$\begin{aligned} \left\| \sum_{i=1}^n \left(\frac{\hat{w}_i^{j,\alpha}}{\sum_l \hat{w}_l^{j,\alpha}} - \frac{w_i^{j,\alpha}}{\sum_l w_l^{j,\alpha}} \right) X_i \right\|_{L^2} &\leq \left\| \frac{\sum_i (\hat{w}_i^{j,\alpha} - w_i^{j,\alpha}) X_i}{\sum_l \hat{w}_l^{j,\alpha}} \right\|_{L^2} \\ &\quad + \left\| \frac{\sum_i w_i^{j,\alpha} X_i \sum_i (\hat{w}_i^{j,\alpha} - w_i^{j,\alpha})}{\sum_l \hat{w}_l^{j,\alpha} \sum_l w_l^{j,\alpha}} \right\|_{L^2}. \end{aligned}$$

To check the order of $g_3 = \sum_i (\hat{w}_i^{j,\alpha} - w_i^{j,\alpha})$ and $g_4 = \sum_i (\hat{w}_i^{j,\alpha} - w_i^{j,\alpha}) X_i$, from (D2),

$$\begin{aligned} |g_3| &\leq \sum_{i=1}^n |\hat{w}_i^{j,\alpha} - w_i^{j,\alpha}| \{\mathbf{I}(\hat{w}_i^{j,\alpha} > 0) + \mathbf{I}(w_i^{j,\alpha} > 0)\} \\ &\leq \frac{C}{nh^{d+1}} \left(\sup_i \|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\| + \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| + \alpha \|(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}} \hat{\mathbf{e}}_j - (\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \right) \\ &\quad \sum_{i=1}^n \{\mathbf{I}(\|\boldsymbol{\vartheta}_i - \boldsymbol{\mu} - \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \leq h) + \mathbf{I}(\|\hat{\boldsymbol{\vartheta}}_i - \hat{\boldsymbol{\mu}} - \alpha(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}} \hat{\mathbf{e}}_j\| \leq h)\}. \end{aligned}$$

Observe $\frac{1}{nh^d} \sum_i \mathbf{I}(\|\boldsymbol{\vartheta}_i - \boldsymbol{\mu} - \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \leq h) \xrightarrow{P} c_1 f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j)$ for some constant $c_1 < \infty$, and that

$$\begin{aligned} \frac{1}{nh^d} \sum_i \mathbf{I}(\|\hat{\boldsymbol{\vartheta}}_i - \hat{\boldsymbol{\mu}} - \alpha(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}} \hat{\mathbf{e}}_j\| \leq h) &\leq \frac{1}{nh^d} \sum_i \{ \mathbf{I}(\|\boldsymbol{\vartheta}_i - \boldsymbol{\mu} - \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \leq 4h) + \\ &\mathbf{I}(\|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\| \geq h) + \mathbf{I}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \geq h) + \mathbf{I}(\alpha\|(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}} \hat{\mathbf{e}}_j - (\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \geq h) \} \end{aligned}$$

converges to $c_2 f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j)$ in probability for some constant $c_2 < \infty$. Here $\frac{1}{nh^d} \sum_i \mathbf{I}(\|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\| \geq h)$, $\frac{1}{h^d} \mathbf{I}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \geq h)$ and $\frac{1}{h^d} \mathbf{I}(\alpha\|(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}} \hat{\mathbf{e}}_j - (\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \geq h)$ are all $o_p(1)$ since

$$\begin{aligned} \mathbb{E}\left\{\frac{1}{nh^d} \sum_i \mathbf{I}(\|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\| \geq h)\right\} &\leq \frac{1}{n} \sum_i \mathbb{E}\|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\| h^{-(d+1)} \rightarrow 0, \\ \mathbb{E}\left\{\frac{1}{h^d} \mathbf{I}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \geq h)\right\} &\leq \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| h^{-(d+1)} \leq \left\{\frac{1}{n} \sum_i \mathbb{E}\|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\| + O\left(\frac{1}{\sqrt{n}}\right)\right\} h^{-(d+1)} \rightarrow 0, \end{aligned}$$

due to the assumptions in Theorem 1. From (48)–(51),

$$\begin{aligned} \mathbb{E}\left\{\frac{1}{h^d} \mathbf{I}(\alpha\|(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}} \hat{\mathbf{e}}_j - (\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \geq h)\right\} &\leq \alpha \mathbb{E}\|(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}} \hat{\mathbf{e}}_j - (\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| h^{-(d+1)} \\ &= O(\mathbb{E}\|\hat{\Sigma}_n - \Sigma\|_F h^{-(d+1)}) = O\left(\mathbb{E}\left\{\frac{1}{n} \sum_i \|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\|^2\right\}^{1/2} + \frac{1}{\sqrt{n}}\right) h^{-(d+1)} \rightarrow 0. \end{aligned}$$

Lemma 2 then implies $|g_3| = O_p\left(\frac{v_n}{h} + \frac{1}{\sqrt{nh}}\right)$. For g_4 , we need to check that

$$\begin{aligned} \frac{1}{nh^d} \sum_i \|X_i\|_{L^2} \mathbf{I}(\|\boldsymbol{\vartheta}_i - \boldsymbol{\mu} - \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \leq h) &\xrightarrow{P} c_1 \|X_{j,\alpha}^{\mathcal{M}}\|_{L^2} f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j), \\ \frac{1}{nh^d} \sum_i \|X_i\|_{L^2} \mathbf{I}(\|\hat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_i\| \geq h) &= o_p(1), \quad \left\{\frac{1}{h^d} \mathbf{I}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \geq h)\right\} \left\{\frac{1}{n} \sum_i \|X_i\|_{L^2}\right\} = o_p(1), \\ \left\{\frac{1}{h^d} \mathbf{I}(\alpha\|(\hat{\lambda}_j^{\mathcal{M}})^{\frac{1}{2}} \hat{\mathbf{e}}_j - (\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j\| \geq h)\right\} \left\{\frac{1}{n} \sum_i \|X_i\|_{L^2}\right\} &= o_p(1). \end{aligned}$$

Using the continuity of ψ^{-1} and f , these terms are analogous to those in g_3 . Thus $\|g_4\|_{L^2} = O_p\left(\frac{v_n}{h} + \frac{1}{\sqrt{nh}}\right)$. Then the second term on the r.h.s. of (52) is seen to be $O_p\left(\frac{v_n}{h} + \frac{1}{\sqrt{nh}}\right)$.

As for the third term on the l.h.s. of (52), let $g_5 = \sum_i w_i^{j,\alpha} (\hat{X}_i^K - X_i)$. Then $\|g_5\|_{L^2} \leq \|\sum_i w_i^{j,\alpha} (\hat{X}_i^K - X_i^K)\|_{L^2} + \|\sum_i \{w_i^{j,\alpha} \sum_{k>K} \xi_{ik} \phi_k\}\|_{L^2}$ and

$$\begin{aligned} \left\|\sum_i \{w_i^{j,\alpha} \sum_{k>K} \xi_{ik} \phi_k\}\right\|_{L^2} &\leq f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j) T^K(X_{j,\alpha}^{\mathcal{M}}) + \\ &\left\|\sum_i \{w_i^{j,\alpha} \sum_{k \geq 1} \xi_{ik} \phi_k\} - f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j) X_{j,\alpha}^{\mathcal{M}}\right\|_{L^2}. \end{aligned}$$

Analogously to g_2 ,

$$\left\| \sum_i \{w_i^{j,\alpha} \sum_{k>K} \xi_{ik} \phi_k\} \right\|_{L^2} = f(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} \mathbf{e}_j) T^K(X_{j,\alpha}^{\mathcal{M}}) + O_p\left(h^2 + \frac{1}{\sqrt{nh^d}}\right). \quad (53)$$

For $\|\sum_i w_i^{j,\alpha}(\hat{X}_i^K - X_i^K)\|_{L^2}$, we use the upper bound for $\|\hat{X}_i^K - X_i^K\|_{L^2}$ provided in (33). From (37) and (47), $\sum_i w_i^{j,\alpha} \|\hat{\mu} - \mu\|_{L^2} = O_p(r_n)$ and $\|\sum_i \{w_i^{j,\alpha} \sum_{k \leq K} \xi_{ik}(\hat{\phi}_k - \phi_k)\}\|_{L^2} \leq \sum_i w_i^{j,\alpha} \|X_i\|_{L^2} \{\sum_{k \leq K} Z_k^{(1)}\}^{1/2} = O_p(K^{\alpha_2 + \frac{1}{2}} r_n)$. From (46), an upper bound for the remaining term in (33) is

$$\begin{aligned} \left\| \sum_{k \leq K} (\hat{\xi}_{ik} - \xi_{ik}) \hat{\phi}_k \right\|_{L^2} &\leq 4\{Q_i^{(1)} \sum_{k \leq K} Z_k^{(1)}\}^{\frac{1}{2}} + 2\sqrt{2}\{Z^{(2)} \sum_{k \leq K} Q_{ik}^{(2)}\}^{\frac{1}{2}} \\ &\quad + 4\{Z^{(2)} \sum_{k \leq K} Z_k^{(1)}\}^{\frac{1}{2}} + 4\tau^* \left\{ \sum_{k \leq K} Z_{ik}^{(3)} \right\}^{\frac{1}{2}} + 2\left\{ \sum_{k \leq K} Q_{ik}^{(3)} \right\}^{\frac{1}{2}} + 2\{Q_i^{(4)} \sum_{k \leq K} Z_k^{(1)}\}^{\frac{1}{2}}. \end{aligned}$$

Here one only needs to pay attention to those terms involving i . Using (41)-(44),

$$\begin{aligned} \sum_{i=1}^n w_i^{j,\alpha} \{Q_i^{(1)}\}^{\frac{1}{2}} &= O_p(1); \quad \sum_{i=1}^n w_i^{j,\alpha} \left\{ \sum_{k \leq K} Q_{ik}^{(2)} \right\}^{\frac{1}{2}} = O_p(K^{\frac{1}{2}}); \quad \sum_{i=1}^n w_i^{j,\alpha} \left\{ \sum_{k \leq K} Z_{ik}^{(3)} \right\}^{\frac{1}{2}} = O_p(K^{\frac{1}{2}\alpha_2 + 1}); \\ \sum_{i=1}^n w_i^{j,\alpha} \left\{ \sum_{k \leq K} Q_{ik}^{(3)} \right\}^{\frac{1}{2}} &= O_p(\sqrt{K} r_n); \quad \sum_{i=1}^n w_i^{j,\alpha} \{Q_i^{(4)}\}^{\frac{1}{2}} = O_p(1). \end{aligned}$$

For the fourth term, observe $\mathbb{E}\{\sum_{k \leq K} Q_{ik}^{(3)}\}^{\frac{1}{2}} \leq \{\mathbb{E} \sum_{k \leq K} Q_{ik}^{(3)}\}^{\frac{1}{2}}$. These rates in conjunction with (39) imply that the leading terms are $4\{Q_i^{(1)} \sum_{k \leq K} Z_k^{(1)}\}^{\frac{1}{2}}$ and $2\{Q_i^{(4)} \sum_{k \leq K} Z_k^{(1)}\}^{\frac{1}{2}}$, resulting in $\|\sum_{k \leq K} (\hat{\xi}_{ik} - \xi_{ik}) \hat{\phi}_k\|_{L^2} = O_p(K^{\alpha_2 + \frac{1}{2}} r_n)$. Thus $\|g_5\|_{L^2} = O_p(T^K(X_{j,\alpha}^{\mathcal{M}}) + h^2 + \frac{1}{\sqrt{nh^d}} + K^{\alpha_2 + \frac{1}{2}} r_n)$, which is also the asymptotic rate of the third term on the r.h.s. of (52).

The last term in (52) is obviously of smaller order, and (30) follows.

Acknowledgements

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

References

ASH, R. B. and GARDNER, M. F. (1975). *Topics in Stochastic Processes*. Academic Press.

- BELKIN, M. and NIYOI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15** 1373–1396.
- BICKEL, P. and LI, B. (2007). Local polynomial regression on unknown manifolds. *Complex Datasets And Inverse Problems: Tomography, Networks And Beyond, ser. IMS Lecture Notes-Monograph Series*. **54** 177186.
- 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.
- COX, T. F. and COX, M. (2001). *Multidimensional Scaling*. Chapman and Hall.
- DE SILVA, V. and TENENBAUM, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems* **15** 721–728.
- DO CARMO, M. P. (1992). *Riemannian Geometry*. Birkhäuser Boston.
- DONOHO, D. L. and GRIMES, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America* **100** 5591–5596.
- DONOHO, D. L. and GRIMES, C. (2005). Image manifolds which are isometric to Euclidean space. *Journal of Mathematical Imaging and Vision* **23** 5–24.
- GASSER, T., MÜLLER, H. G., KÖHLER, W., MOLINARI, L. and PRADER, A. (1984). Non-parametric regression analysis of growth curves. *Annals of Statistics* **12** 210–229.
- GERVINI, D. and GASSER, T. (2004). Self-modelling warping functions. *Journal of the Royal Statistical Society Series B* **66** 959–971.
- GERVINI, D. and GASSER, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92** 801–820.

- GRENANDER, U. (1950). Stochastic processes and statistical inference. *Arkiv för Matematik* **1** 195–277.
- HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics* **35** 70–91.
- HELGASON, S. (2001). *Differential Geometry, Lie Groups, and Symmetric Spaces*. American Mathematical Society.
- HUCKEMANN, S. (2011). Inference on 3d Procrustes means: Tree bole growth, rank deficient diffusion tensors and perturbation models. *Scandinavian Journal of Statistics* **38** 1467–9469.
- IZEM, R. and MARRON, J. (2007). Analysis of nonlinear modes of variation for functional data. *Electronic Journal of Statistics* **1** 641–676.
- JONES, M. C. and RICE, J. A. (1992). Displaying the important features of large collections of similar curves. *American Statistician* **46** 140–145.
- KATO, T. (1966). *Perturbation Theory for Linear Operators*. Springer.
- KENDALL, D. G., BARDEN, D., CARNE, T. K. and LE, H. (1999). *Shape and Shape Theory*. Wiley N.Y.
- KNEIP, A. and GASSER, T. (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics* **20** 1266–1305.
- KNEIP, A. and UTIKAL, K. (2001). Inference for density families using functional principal components analysis. *Journal of the American Statistical Association* **96** 519–531.
- LENG, X. and MÜLLER, H. G. (2006). Time ordering of gene co-expression. *Biostatistics* **7** 569–584.
- MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press Inc.

- MÜLLER, H.-G. and YAO, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103** 1534–1544.
- RIESZ, F. and SZ-NAGY, B. (1990). *Functional Analysis*. Dover Publications.
- ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323—2326.
- SPELLMAN, P., SHERLOCK, G. and ZHANG, M. (1998). Comprehensive identification of cell cycle-regulated gene of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9** 3273–3297.
- TANG, R. and MÜLLER, H. G. (2009). Time-synchronized clustering of gene expression trajectories. *Biostatistics* **10** 32–45.
- TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.
- TUDDENHAM, R. and SNYDER, M. (1954). Physical growth of California boys and girls from birth to age 18. *Calif. Publ. Child Deve.* **1** 183–364.
- WANG, K. and GASSER, T. (1999). Synchronizing sample curves nonparametrically. *Annals of Statistics* **27** 439–460.
- 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.

Table 1: Fraction of distances explained $FDE = 1 - \frac{\|\hat{D}^p - D\|_F}{\|D\|_F}$ defined in (24) for isometric manifold fits with different dimension d (other parameters are optimized), for two signal-to-noise ratios R and manifolds $\mathcal{M}_1 - \mathcal{M}_3$.

Manifold	R	d				
		1	2	3	4	5
\mathcal{M}_1	0.1	.998	.999	.999	.999	.999
	0.5	.9778	.993	.995	.995	.996
\mathcal{M}_2	0.1	.914	.988	.994	.996	.996
	0.5	.902	.971	.974	.978	.980
\mathcal{M}_3	0.1	.699	.932	.957	.977	.980
	0.5	.639	.906	.948	.955	.958
Growth		.947	.972	.980	.985	.988
Yeast		.891	.949	.981	.983	.984
Mortality		.878	.954	.973	.980	.982

Table 2: Simulation results for manifolds $\mathcal{M}_1 - \mathcal{M}_3$. Each Monte Carlo is based on a sample of 200 curves, sampled at 30 equidistant time points on $[-4, 4]$. The values in the table are mean squared prediction errors, $\text{MSPE} = \frac{1}{200} \sum_{i=1}^{200} \|X_i - \hat{X}_i\|_{L^2}^2$, where $\hat{X}_i = \hat{X}_i^L$, the prediction using functional principal component analysis, or $\hat{X}_i^{\mathcal{M}}$, using the proposed manifold method, and relative squared prediction error, $\text{RSPE} = \frac{\sum_{i=1}^{200} \|X_i - \hat{X}_i\|_{L^2}^2}{\sum_{i=1}^{200} \|X_i - \bar{X}\|_{L^2}^2}$, where $\bar{X} = \frac{1}{200} \sum_{i=1}^{200} X_i$. The dimension L is the number of included FPCs for \hat{X}_i^L , and d is the intrinsic dimension assumed for $\hat{X}_i^{\mathcal{M}}$. Comparisons are made at the same dimensions for signal-to-noise ratios $R = 0.1$ and $R = 0.5$. Results are mean values based on 100 Monte Carlo runs.

Manifold	R	Method	MSPE with different L or d					RSPE with different L or d (%)				
			1	2	3	4	5	1	2	3	4	5
\mathcal{M}_1	0.1	\hat{X}_i^L	.159	.034	.025	.021	.021	41.4	9.9	6.4	5.5	5.5
		$\hat{X}_i^{\mathcal{M}}$.027	.015	.015	.014	.015	7.0	3.9	3.9	3.7	3.8
	0.5	\hat{X}_i^L	.173	.061	.057	.058	.058	45.1	15.7	14.8	15.0	15.1
		$\hat{X}_i^{\mathcal{M}}$.090	.046	.046	.049	.053	23.3	12.0	11.8	12.7	13.8
\mathcal{M}_2	0.1	\hat{X}_i^L	.054	.022	.013	.008	.007	44.1	17.6	10.3	6.7	5.7
		$\hat{X}_i^{\mathcal{M}}$.022	.009	.007	.006	.006	18.0	7.5	5.4	4.6	4.6
	0.5	\hat{X}_i^L	.055	.025	.019	.018	.018	44.9	20.4	15.9	14.4	14.4
		$\hat{X}_i^{\mathcal{M}}$.030	.017	.015	.014	.014	24.8	13.8	12.0	11.5	11.6
\mathcal{M}_3	0.1	\hat{X}_i^L	.148	.059	.031	.023	.020	58.7	23.3	12.5	9.1	7.7
		$\hat{X}_i^{\mathcal{M}}$.088	.025	.020	.020	.019	34.9	10.0	8.1	7.7	7.6
	0.5	\hat{X}_i^L	.154	.071	.053	.048	.048	61.2	28.3	21.1	19.0	19.0
		$\hat{X}_i^{\mathcal{M}}$.124	.059	.047	.045	.044	49.0	23.6	18.6	17.7	17.6

Table 3: Simulation results for manifolds $\mathcal{M}_1 - \mathcal{M}_3$, demonstrating mean ratios of mean squared prediction error (MSPE) for P-ISOMAP over MSPE of ISOMAP when d equals the intrinsic dimension, for two signal-to-noise ratios R .

R	\mathcal{M}_1	\mathcal{M}_2	\mathcal{M}_3
0.1	.9676	.9679	.9402
0.5	.8121	.8879	.8302

Table 4: The mean squared prediction errors using different values for ϵ (for dimensions $d = 1$ for \mathcal{M}_1 and 2 for $\mathcal{M}_2, \mathcal{M}_3$, while the other parameters are optimized). Note that when ϵ is small, some sample points will be treated as outliers as they are not connected to other sample points, which renders the MSPE for small ϵ inaccurate.

Manifold	R	ϵ				
		3	5	8	12	16
\mathcal{M}_1	0.1	.029	.027	.031	.029	.033
	0.5	.116	.102	.090	.129	.135
\mathcal{M}_2	0.1	.008	.010	.009	.010	.010
	0.5	.020	.018	.018	.017	.017
\mathcal{M}_3	0.1	.029	.040	.025	0.27	.033
	0.5	.052	.059	.065	.059	.066

Table 5: Mean squared prediction errors (MSPE) and relative squared prediction errors (RSPE) (in percent) for the three application data sets in section 6.

Data	Method	MSPE with different L or d					RSPE with different 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	61.9	46.7	49.9	49.6	45.9
	\hat{X}_i^M	10.7	9.46	9.06	9.21	9.08	38.7	34.2	32.8	33.3	32.9
Yeast	\hat{X}_i^L	.639	.382	.257	.205	.203	66.9	40.0	26.9	21.5	21.3
	\hat{X}_i^M	.468	.278	.231	.210	.206	49.0	29.1	24.2	22.0	21.6
Mortality	\hat{X}_i^L	7.38	6.34	5.44	5.48	5.21	54.4	47.3	40.0	40.3	38.3
	\hat{X}_i^M	6.77	5.64	5.40	5.26	4.98	49.8	41.3	39.7	38.7	36.5

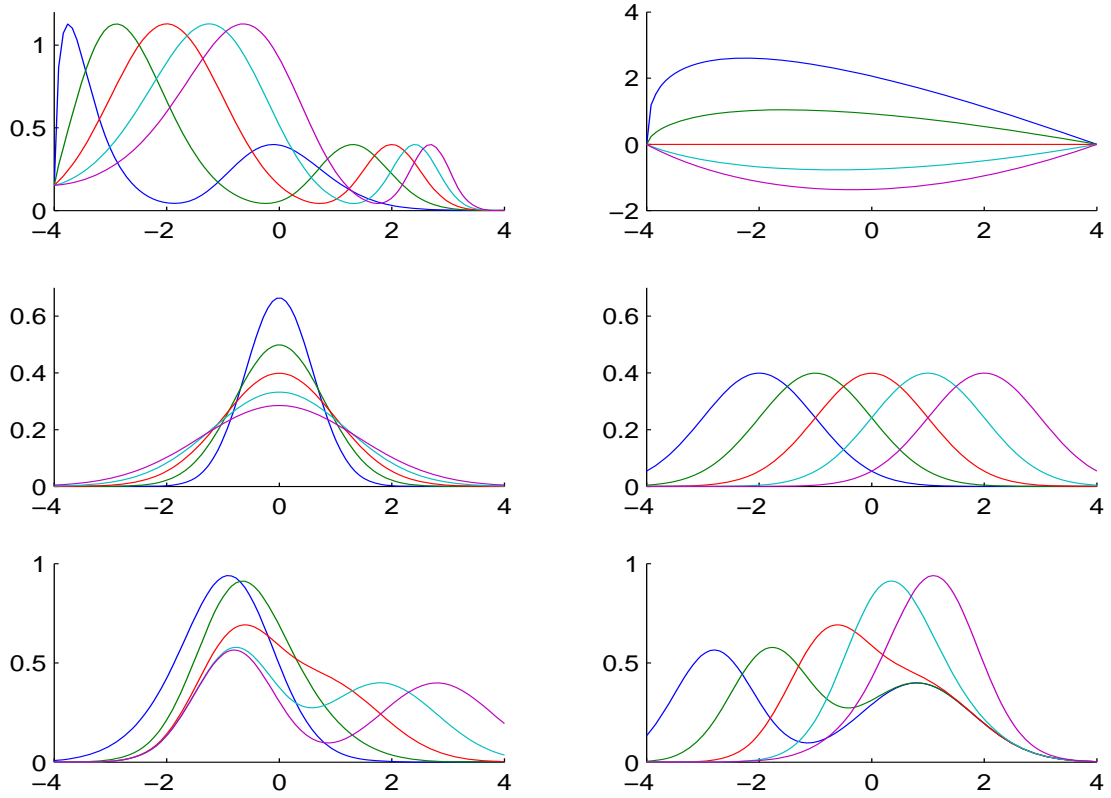


Figure 1: Manifolds $\mathcal{M}_1 - \mathcal{M}_3$. Top left panel: Functions on \mathcal{M}_1 for $\alpha = 0.6, 0.8, 1.0, 1.2, 1.4$. Top right panel: Corresponding identity-subtracted warping functions $h_\alpha(t) - t$. Middle left panel: Functions on \mathcal{M}_2 for $\alpha = 0.4, 0.7, 1.0, 1.3,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

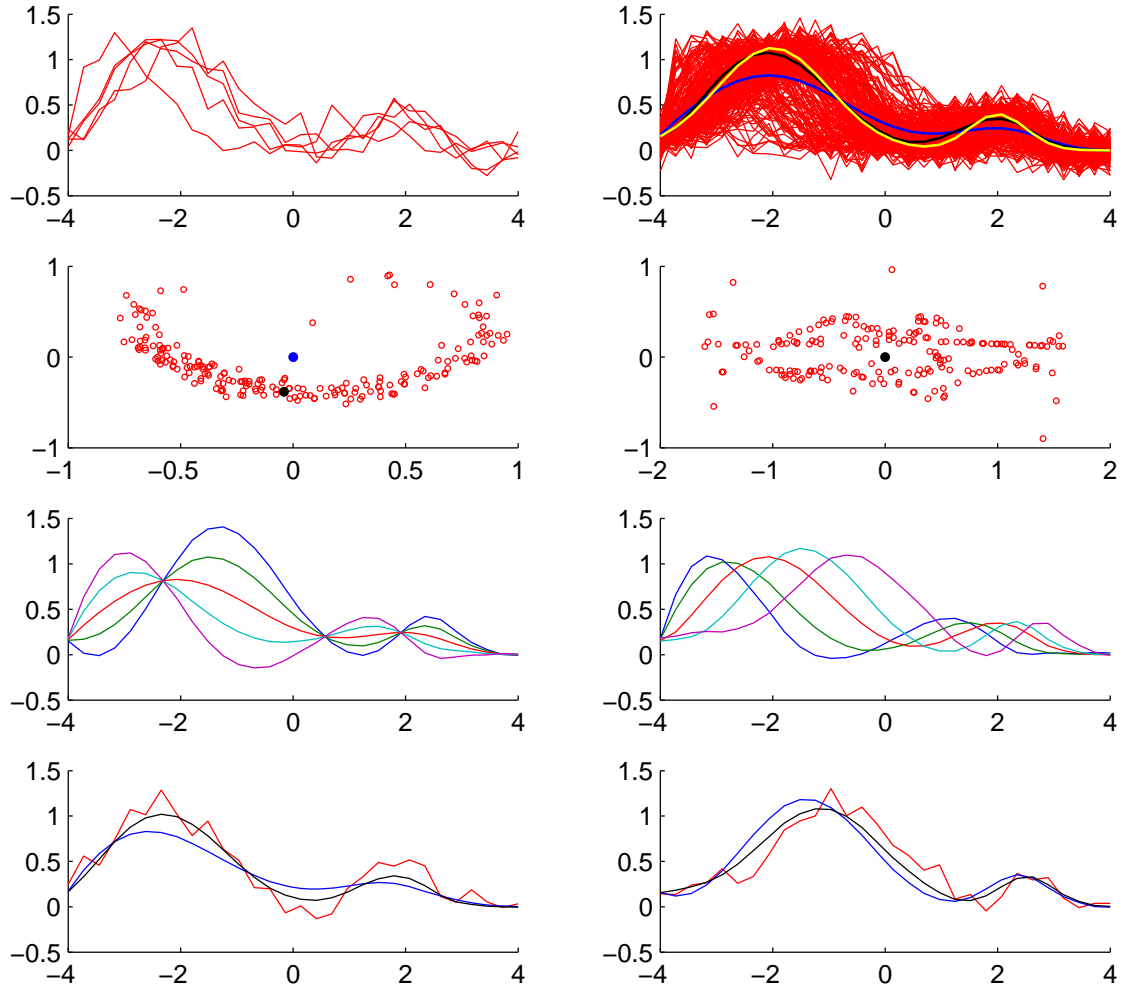


Figure 2: Simulated data for manifold \mathcal{M}_1 . Top left panel: Five randomly selected curves. Top right panel: Common shape function (yellow, corresponds to target mean), estimated manifold mean $\hat{\mu}^{\mathcal{M}}$ (21) (black) and the L^2 mean (blue). Second row: Scatter plot of second 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 bold blue dot represents the L^2 mean. Third row: Estimates of principal component based mode $X_{1,\alpha}$ (7) (left) and of manifold mode $X_{1,\alpha}^{\mathcal{M}}$ (9) (right) of functional variation for $\alpha = -2, -1, 0, 1, 2$. Bottom row: Two randomly selected curves (red), with the corresponding principal component based predictions \hat{X}_i^L (12) (blue), and manifold based predictions $\hat{X}_i^{\mathcal{M}}$ (20) (black) for $L = d = 2$.

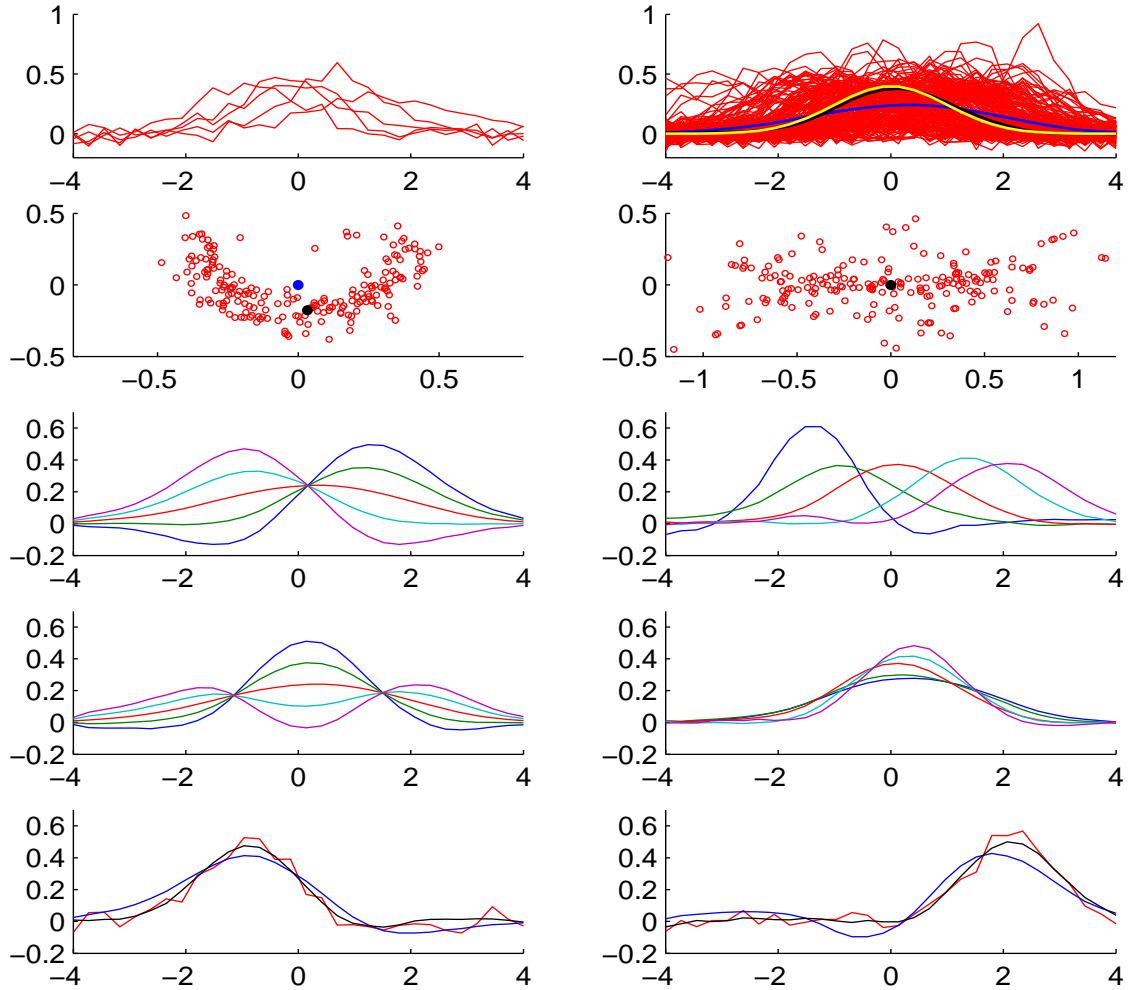


Figure 3: Simulated data for manifold \mathcal{M}_2 . Top left panel: Five randomly selected curves. Top right panel: Standard Gaussian density (yellow, corresponds to target mean), estimated manifold mean $\hat{\mu}^{\mathcal{M}}$ (21) (black) and the L^2 mean (blue). Second row: Scatter plot of second versus first FPC (left) and second versus first FMC (right), where the bold black dot represents the manifold mean and the bold blue dot represents the L^2 mean. Third row: Estimates of principal component based mode $X_{1,\alpha}$ (7) (left) and of manifold mode $X_{1,\alpha}^{\mathcal{M}}$ (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 (red), with the corresponding principal component based predictions \hat{X}_i^L (12) (blue), and manifold based predictions $\hat{X}_i^{\mathcal{M}}$ (20) (black) for $L = d = 3$.

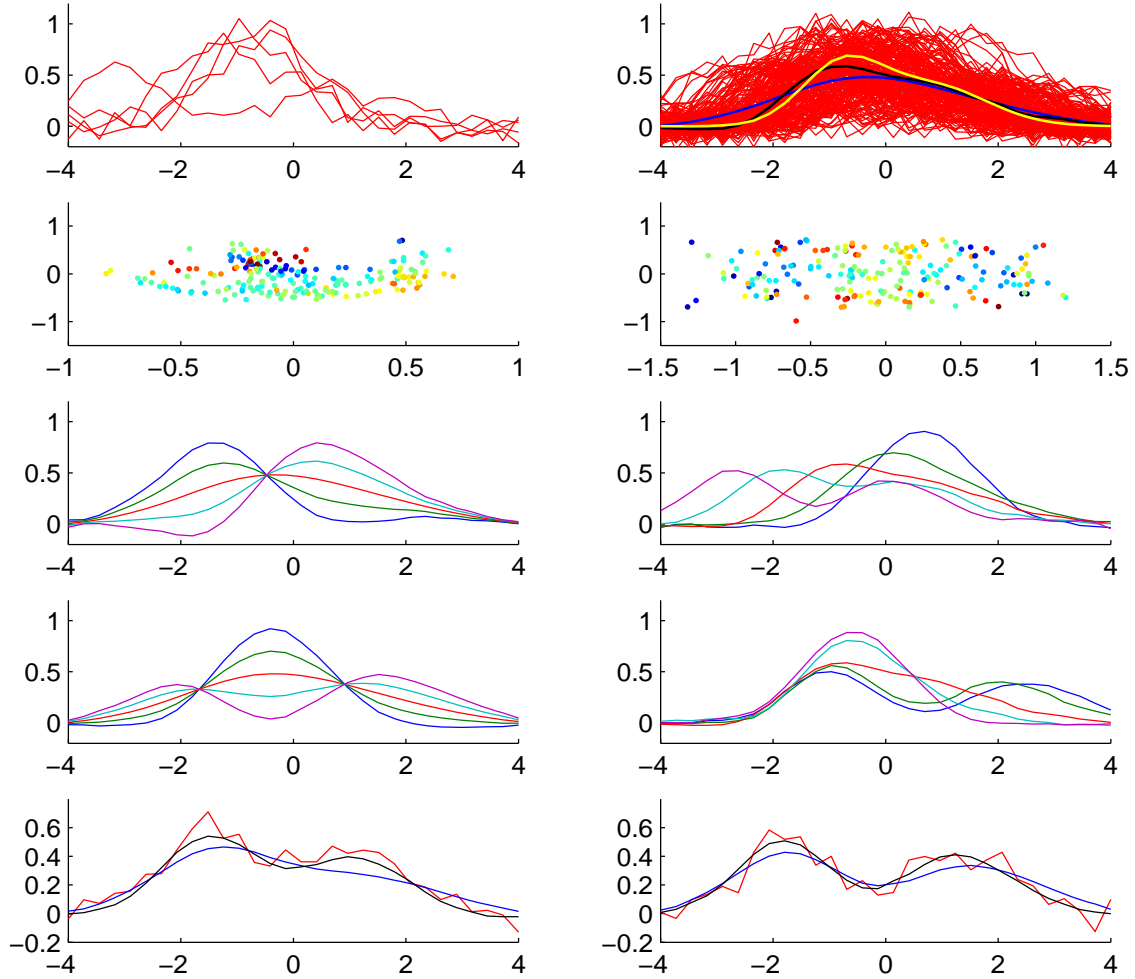


Figure 4: Simulated data for manifold \mathcal{M}_3 . Top left panel: Five randomly selected curves. Top right panel: Curve with no time shifts (yellow, corresponds to target mean), estimated manifold mean $\hat{\mu}^{\mathcal{M}}$ (21) (black) and the L^2 mean (blue). Second row: Contour scatter plot 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 $X_{1,\alpha}$ (7) (left) and of manifold mode $X_{1,\alpha}^{\mathcal{M}}$ (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 (red), with the corresponding principal component based predictions \hat{X}_i^L (12) (blue), and manifold based predictions $\hat{X}_i^{\mathcal{M}}$ (20) (black) for $L = d = 3$.

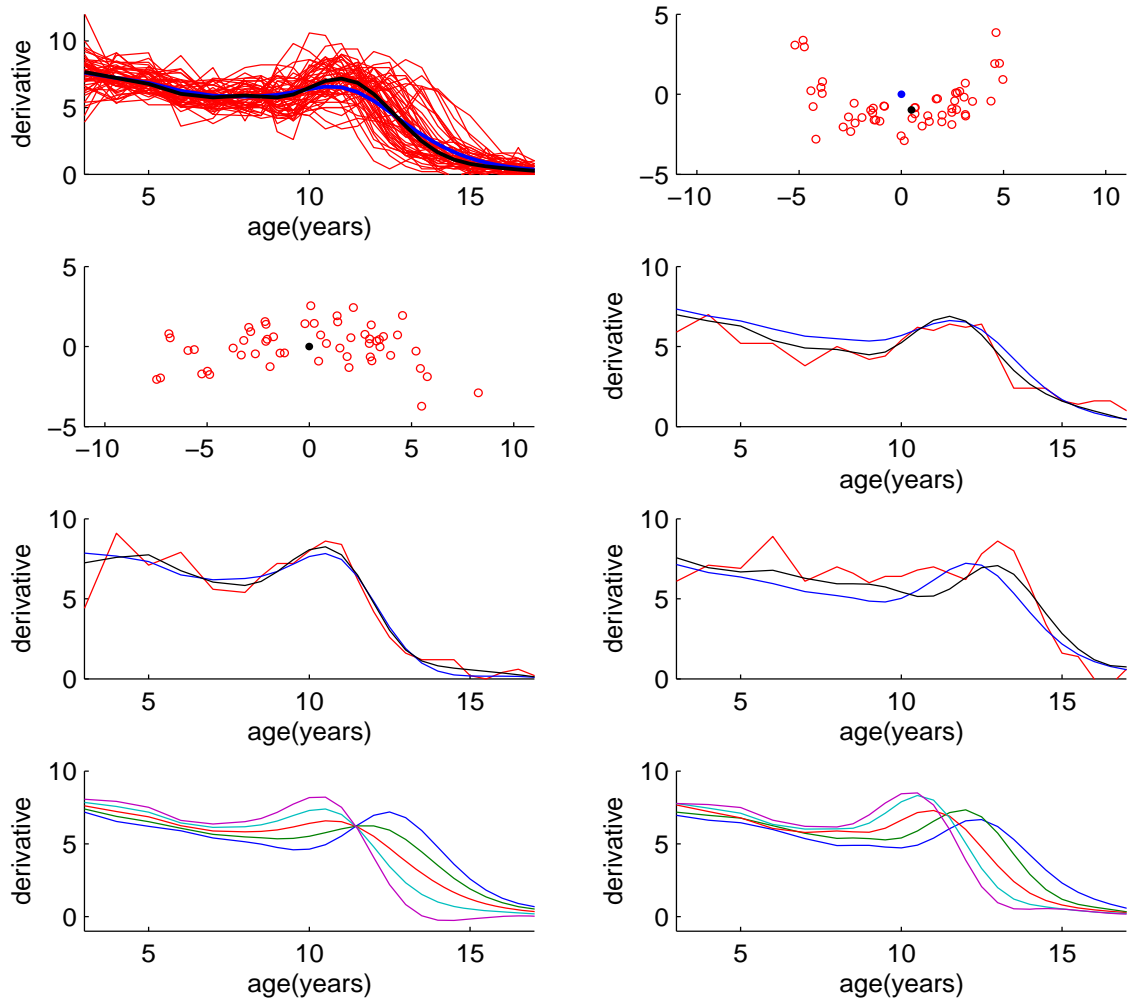


Figure 5: Berkeley growth data for girls. Top left panel: Derivatives with the cross-sectional mean (blue) and estimated manifold mean $\hat{\mu}^{\mathcal{M}}$ (21) (black). Top right panel: Scatter plot of second versus first FPC, where the bold black dot represents the manifold mean and the bold blue dot 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 (red), with the corresponding principal component based predictions \hat{X}_i^L (12) (blue), and manifold based predictions $\hat{X}_i^{\mathcal{M}}$ (20) (black) for $L = d = 2$. Bottom panels: Estimates of principal component based mode $X_{1,\alpha}$ (7) (left) and of manifold mode $X_{1,\alpha}^{\mathcal{M}}$ (9) (right) of functional variation for $\alpha = -2, -1, 0, 1, 2$.

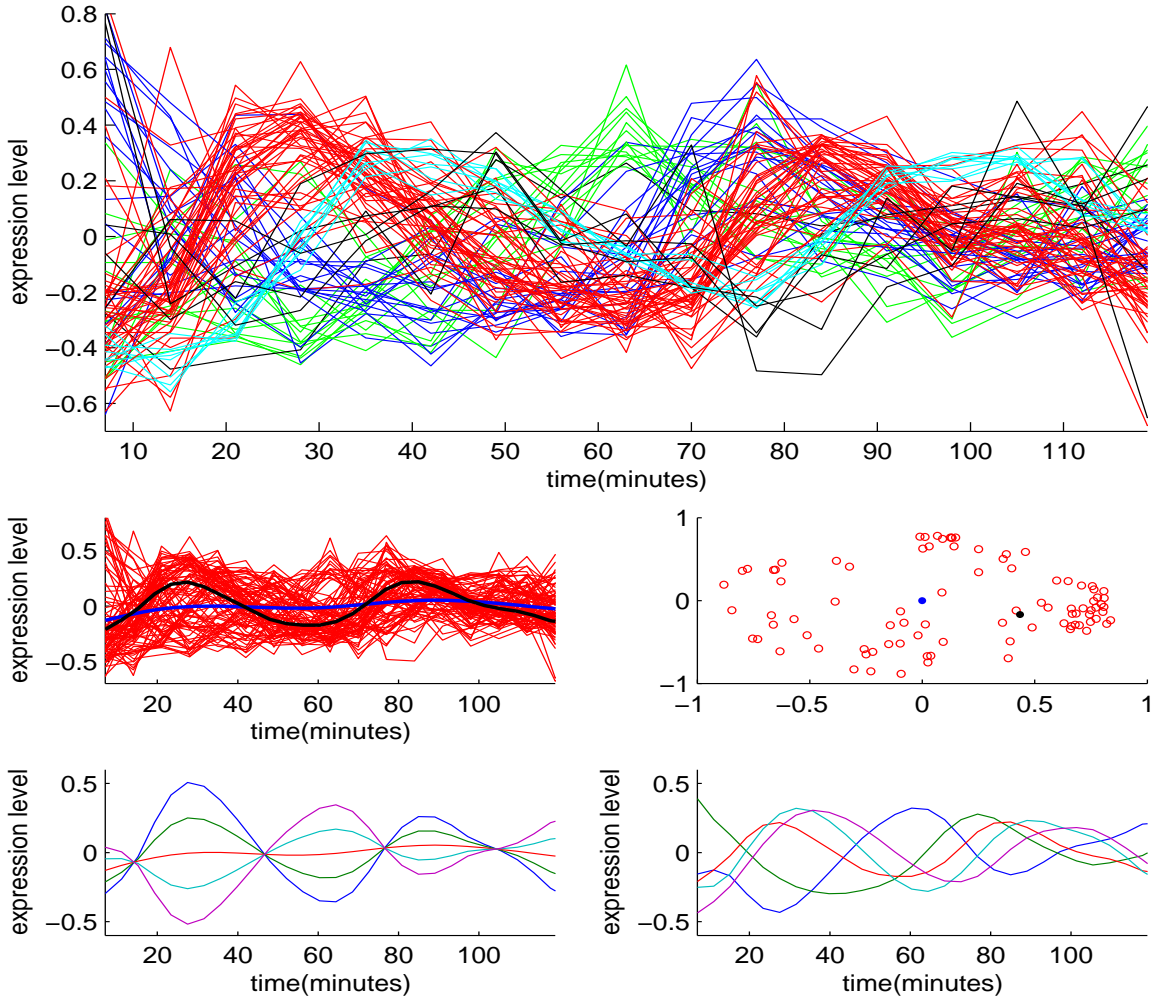


Figure 6: Yeast cell cycle gene expression data. Top panel: All trajectories in different colors according to cluster membership: G1 (red), S (cyan), G2/M (green), M/G1 (blue) and S/G2 (black). Middle left panel: Estimated manifold mean $\hat{\mu}^{\mathcal{M}}$ (21) (black) and cross-sectional mean (blue). Middle right panel: Scatter plot of second versus first FPC, where the bold blue dot indicates cross-sectional mean and the bold black dot indicates manifold mean. Bottom panels: Estimates of principal component based mode $X_{1,\alpha}$ (7) (left) and of manifold mode $X_{1,\alpha}^{\mathcal{M}}$ (9) (right) of functional variation for $\alpha = -2, -1, 0, 1, 2$.

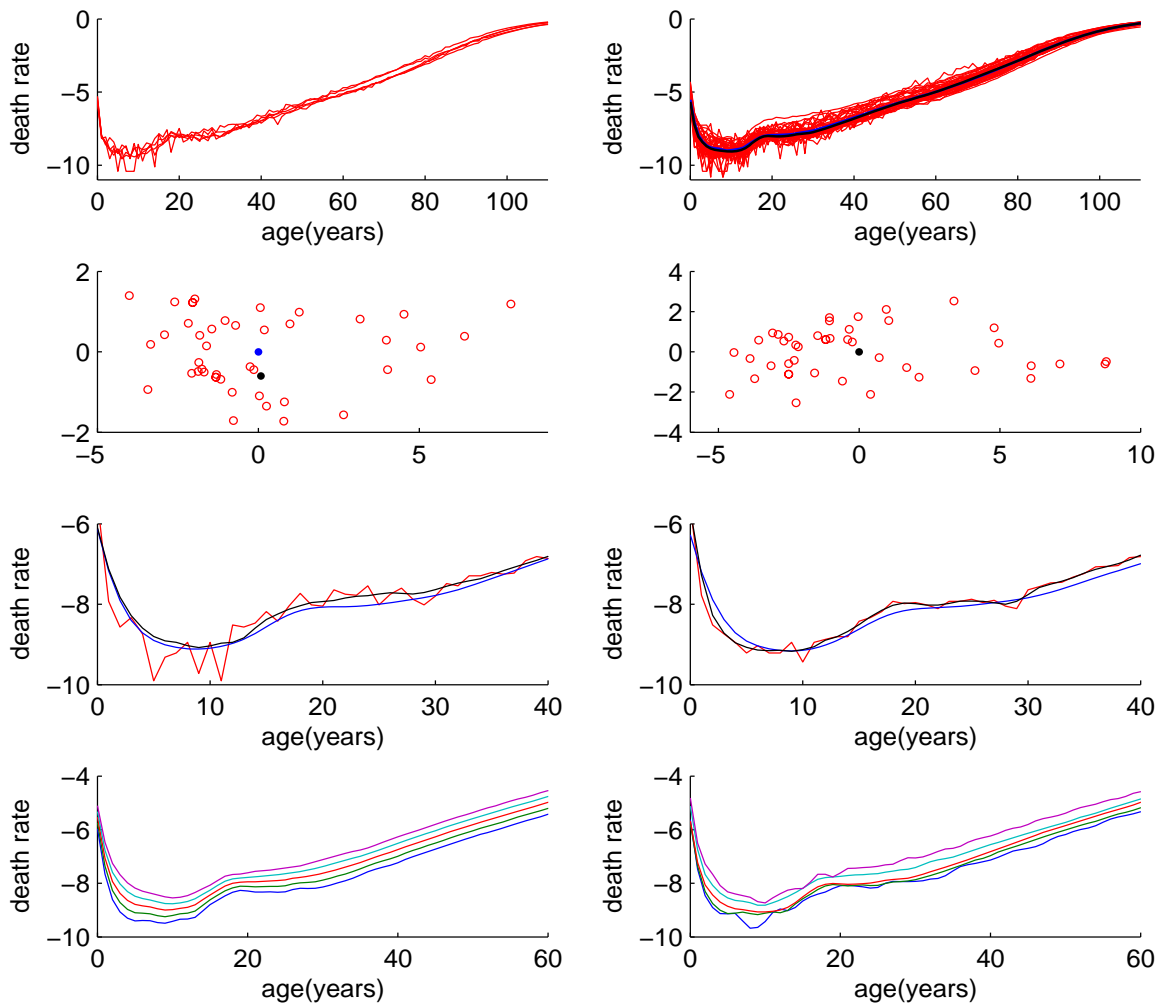


Figure 7: Human mortality data. Top left panel: Death rates for five randomly selected countries. Top right panel: Estimates of cross-sectional mean (blue) and manifold mean $\hat{\mu}^{\mathcal{M}}$ (13) (black). Second row: Scatter plots of second versus first FPC (left) and second versus first FMC (right), where the bold blue dot indicates the cross-sectional mean and the bold black dot indicates the manifold mean. Third row: Two randomly selected curves (red), with the corresponding principal component based predictions \hat{X}_i^L (12) (blue), and manifold based predictions $\hat{X}_i^{\mathcal{M}}$ (20) (black) for $L = d = 3$. Bottom panels: Estimates of principal component based mode $X_{1,\alpha}$ (7) (left) and of manifold mode $X_{1,\alpha}^{\mathcal{M}}$ (9) (right) of functional variation for $\alpha = -2, -1, 0, 1, 2$.