Mixture Inner Product Spaces and Their Application to Functional Data Analysis

Zhenhua Lin[1], Hans-Georg Müller[2] and Fang Yao[1,3]

Abstract

We introduce the concept of mixture inner product spaces associated with a given separable Hilbert space, which feature an infinite-dimensional mixture of finite-dimensional vector spaces and are dense in the underlying Hilbert space. Any Hilbert valued random element can be arbitrarily closely approximated by mixture inner product space valued random elements. While this concept can be applied to data in any infinite-dimensional Hilbert space, the case of functional data that are random elements in the L^2 space of square integrable functions is of special interest. For functional data, mixture inner product spaces provide a new perspective, where each realization of the underlying stochastic process falls into one of the component spaces and is represented by a finite number of basis functions, the number of which corresponds to the dimension of the component space. In the mixture representation of functional data, the number of included mixture components used to represent a given random element in L^2 is specifically adapted to each random trajectory and may be arbitrarily large. Key benefits of this novel approach are, first, that it provides a new perspective on the construction of a probability density in function space under mild regularity conditions, and second, that individual trajectories possess a trajectory-specific dimension that corresponds to a latent random variable, making it possible to use a larger number of components for less smooth and a smaller number for smoother trajectories. This enables flexible and parsimonious modeling of heterogeneous trajectory shapes. We establish estimation consistency of the functional mixture density and introduce an algorithm for fitting the functional mixture model based on a modified expectation-maximization algorithm. Simulations confirm that in comparison to traditional functional principal component analysis the proposed method achieves similar or better data recovery while using fewer components on average. Its practical merits are also demonstrated in an analysis of egg-laying trajectories for medflies.

Key words and phrases: Basis; Density; Functional Data Analysis; Infinite Mixture; Trajectory Representation.

AMS Subject Classification: 62G05, 62G08

[1] Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada

[2] Department of Statistics, University of California, One Shields Avenue, Davis, California 95616, U.S.A.

[3] Corresponding author, email: fyao@utstat.toronto.edu.

1 Introduction

Introducing the concept of mixture inner product spaces is motivated by one of the basic problems in functional data analysis, namely to efficiently represent functional trajectories by dimension reduction. Functional data correspond to random samples $\{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n\}$ drawn from a square-integrable random process defined on a finite interval D, $\tilde{X} \in L^2(D)$. Random functions $\tilde{X}_i$ are generally considered to be inherently infinite-dimensional and therefore finite-dimensional representations are essential. A commonly employed approach for dimension reduction is to expand the functional data in a suitable basis in function space and then to represent the random functions in terms of the sequence of expansion coefficients. This approach has been very successful and has been implemented with B-spline bases (Ramsay and Silverman, 2005) and eigenbases, which consist of the eigenfunctions of the covariance operator of the underlying stochastic process that generates the data. The estimated eigenbasis expansion then gives rise to functional principal component analysis, which was introduced in a rudimentary form in Rao (1958) for the analysis of growth curves. Earlier work on eigendecompositions of square integrable stochastic processes (Grenander, 1950; Gikhman and Skorokhod, 1969) paved the way for statistical approaches.

By now there is a substantial literature on functional principal component analysis, including basic developments (Besse and Ramsay, 1986; Castro *et al.*, 1986), advanced smoothing implementations and the concept of modes of variation (Rice and Silverman, 1991; Silverman, 1996), theory (Boente and Fraiman, 2000; Kneip and Utikal, 2001; Hall and Hosseini-Nasab, 2006), as well as a unified framework that covers functional principal component analysis for functional data with both sparse and dense designs and therefore brings many longitudinal data under this umbrella (Yao *et al.*, 2005; Li and Hsing, 2010; Zhang and Wang, 2016). One of the attractions of functional principal component analysis is that for any number of included components the resulting finite-dimensional approximation to the infinite-dimensional process explains most of the variation. This has contributed to the enduring popularity of functional principal component analysis (Li and Guan, 2014; Chen and Lei, 2015), which differs in essential ways from classical multivariate principal component analysis, due to the smoothness and infinite dimensionality of the functional objects.

Existing methods assume a common structural dimension for this approximation (Hall and Vial, 2006; Li *et al.*, 2013), where for asymptotic consistency it is assumed that the number of included components, which is the same for all trajectories in the sample, increases with sample size to ensure asymptotic unbiasedness. To determine an adequate number of components based on observed functional data that is applied across the sample to approximate the underlying processes reasonably well is crucial for the application of functional principal component analysis. This is challenging for applications in which the trajectories recorded for different subjects exhibit different levels of complexity. We introduce here an alternative to the prevailing paradigm that the observed functional data are all infinite-dimensional objects, which are then approximated through a one-size-fits-all sequence of increasingly complex approximations. The proposed alternative model is to assume that each observed random trajectory is composed of only finitely many components, where the number of components that constitutes an observed trajectory may be arbitrarily large without upper bound and varies across the observed trajectories. This means that while each trajectory can be fully represented without residual by its projections on a finite number of components, the overall process is still infinite-dimensional as no finite dimension suffices to represent it: For each fixed dimension d, there generally exist trajectories that require more than d components for adequate representation. A key feature of this new model is that the number of components used to represent a trajectory depends on the trajectory to be represented.

In this paper, we develop the details of this model and show in data analysis and simulations that its implementation leads to more parsimonious representations of heterogeneous functional data when compared with classical functional principal component analysis. Its relevance for functional data analysis motivates us to develop this model in the context of a general infinite-dimensional separable Hilbert space; we note that all Hilbert spaces considered in this paper are assumed to be separable. For any given infinite-dimensional Hilbert space and an orthonormal basis of this space, we construct an associated *mixture inner product space* (MIPS). The mixture inner product space consists of an infinite mixture of vector spaces with different dimensions $d, d = 1, 2, 3, \ldots$. We investigate properties of probability measures on these dimension mixture spaces and show that the mixture inner product space associated with a given Hilbert space is dense in the Hilbert space and is well suited to approximate individual Hilbert space elements as well as probability measures on the Hilbert space.

The mixture inner product space concept has direct applications in functional data analysis. It is intrinsically linked to a trajectory-adaptive choice of the number of included components and moreover can be harnessed to construct a density for functional data. The density problem when viewed in the Hilbert space L^2 arises due to the well-known non-existence of a probability density for functional data with respect to Lebesgue measure in L^2, which is a consequence of the low small ball probabilities (Li and Linde, 1999; Niang, 2002) in this space. The lack of a density is a drawback that negatively impacts various methods of functional data analysis. For example, it is difficult to rigorously define modes, likelihoods or other density-dependent methods, such as functional clustering or functional Bayes classifiers. It has therefore been proposed to approach this problem by defining a sequence of approximating densities, where one considers the joint density of the first K functional principal components, as K increases slowly with sample size. This leads to a sequence of finite-dimensional densities that can be thought of as a surrogate density (Delaigle and Hall, 2010; Bongiorno and Goia, 2016). This approach bypasses but does not resolve the key issue that a density in the functional space L^2 does not exist.

In contrast, if the random functions lie in a mixture inner product space, which includes functions of arbitrarily large dimension, one can construct a well-defined target density by introducing a suitable measure for mixture distributions. This density is a mixture of densities on vector spaces of various dimensions d and its existence follows from the fact that a density exists with respect to the usual Lebesgue measure for each component space, which is a finite-dimensional vector space. Therefore, the proposed mixture inner product space approach is of relevance for the foundations of the theory of functional data analysis.

The paper is organized as follows. We develop the concept of mixture inner product spaces and associated probability measures on such spaces in Section 2 and then apply it to functional data analysis in Section 3. This is followed by simulation studies in Section 4 and an application of the proposed method to a real data set in Section 5. Conclusions are in Section 6. All proofs and technical details are in the Appendix.

2 Random Elements in Mixture Inner Product Spaces

In the theory of functional data analysis, functional data can be alternatively viewed as random elements in L^2 or as realizations of stochastic processes. Under joint measurability assumptions, these perspectives coincide; see Chapter 7 of Hsing and Eubank (2015). We adopt the random element perspective in this paper, which is more convenient as we will develop the concept of a mixture inner product space (MIPS) first for general infinite-dimensional Hilbert spaces, and will then take up the special case of functional data and L^2 in Section 3. In this section we consider probability measures on Hilbert spaces and random elements that are Hilbert space valued random variables.

2.1 Mixture Inner Product Spaces

Let H be an infinite-dimensional Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and induced norm $\|\cdot\|$. Let $\Phi = (\phi_1, \phi_2, \ldots)$ be a complete orthonormal basis (CONS) of H. We also assume that the ordering of the sequence $\phi_1, \phi_2, \ldots$ is given and fixed. Define H_k, $k = 0, 1, \ldots,$ as the linear subspace spanned by $\phi_1, \phi_2, \ldots, \phi_k$, where $H_0 = \emptyset$, and set $S_k = H_k \backslash H_{k-1}$ for $k = 1, 2, \ldots$ and $S = \bigcup_{k=1}^{\infty} S_k$, where also $S = \bigcup_{k=1}^{\infty} H_k$. Then S is an infinite-dimensional linear subspace of H with inner product inherited from H. Since S has an inner product and is a union of the k-dimensional subsets S_k, we refer to S as mixture inner product space (MIPS). The definition of S_k depends on Φ and thus on the ordered sequence $\phi_1, \phi_2, \ldots$, while S depends on Φ only in the sense that any permutation of $\phi_1, \phi_2, \ldots$ yields the same space $S = S(\Phi)$. It is easy to see that two CONS $\Phi = (\phi_1, \phi_2, \ldots)$ and $\Psi = (\psi_1, \psi_2, \ldots)$ result in the same MIPS, i.e., $S(\Phi) = S(\Psi)$, if and only if for each $k = 1, 2, \ldots$, there exists a positive integer $n_k < \infty$, positive integers $k_1, k_2, \ldots, k_{n_k} < \infty$ and real numbers $a_{k_1}, a_{k_2}, \ldots, a_{k_{n_k}}$, such that $\phi_k = \sum_{j=1}^{n_k} a_{k_j} \psi_{k_j}$.

In the sequel, we assume a CONS Φ is pre-determined, and $S(\Phi)$ is simply denoted by S. Let $\mathscr{B}(H)$ be the Borel σ-algebra of H and $(\Omega, \mathscr{E}, P)$ a probability space. A H-valued random element X_H is a $\mathscr{E}$-$\mathscr{B}(H)$ measurable mapping from Ω to H. Recall that S is an inner product space, and hence it has its own Borel σ-algebra $\mathscr{B}(S)$. Therefore, S-valued random elements can be defined as $\mathscr{E}$-$\mathscr{B}(S)$ measurable maps from Ω to S. The following proposition establishes some basic properties of MIPS, where it should be noted that S is a proper subspace of H; for example, $h = \sum_{k=1}^{\infty} 2^{-k} \phi_k$ is in H but not S.

Proposition 1. *Let S be a MIPS of H. Then,*

1. *S is a dense subset of H;*
2. *$S \in \mathscr{B}(H)$ and $\mathscr{B}(S) \subset \mathscr{B}(H)$;*
3. *Every S-valued random element X_S is also an H-valued random element.*

An important consequence of the denseness of S is that any H-valued random element can be uniformly approximated by S-valued random elements to an arbitrary precision: Consider $\xi_j = \langle X, \phi_j \rangle$ and $X_k = \sum_{j=1}^{k} \xi_j \phi_j$. For each $j, k = 1, 2, \ldots$, define $\Omega_{j,k} = \{\omega \in \Omega : \|X - X_k\|_H < j^{-1}\} \backslash \Omega_{j,k-1}$, with $\Omega_{1,0} = \emptyset$. Because $\|X(\omega) - X_k(\omega)\|_H \to 0$ for each $\omega \in \Omega$, $\Omega_{j,1}, \Omega_{j,2}, \ldots$ form a measurable partition of Ω for each j. Defining $Y_j(\omega) = \sum_{k=1}^{\infty} X_k(\omega) 1_{\omega \in \Omega_{j,k}}$, where $1_{\omega \in \Omega_{j,k}}$ is the indicator function of $\Omega_{j,k}$, for each ω, there is a k such that $Y_j(\omega) = X_k(\omega) \in S$. Moreover, if $A \in \mathscr{B}(S)$, then $Y_j^{-1}(A) = \bigcup_{k=1}^{\infty} (X_k^{-1}(A) \cap \Omega_{j,k}) \in \mathscr{E}$, as each X_k is measurable. Therefore, each Y_j is $\mathscr{E}$-$\mathscr{B}(S)$ measurable and hence an S-valued random element. Finally, the construction of Y_j guarantees that $\sup_{\omega \in \Omega} \|X(\omega) - Y_j(\omega)\|_H < j^{-1} \to 0$ as $j \to \infty$. This leads to the following uniform approximation result.

Theorem 1. *If X is a H-valued random element and S is a MIPS of H, there exists a sequence of S-valued random elements $Y_1, Y_2, \ldots$, such that* $\sup_{\omega \in \Omega} \|X(\omega) - Y_j(\omega)\|_H \to 0$ *as $j \to \infty$.*

From the above discussion, we see that in approximating X with precision j^{-1}, the number of components used for different ω might be different. For example, if $\omega \in \Omega_{j,k}$, then k components are used. This adaptivity of S-valued random elements can lead to an overall more parsimonious approximation of X compared to approximations with fixed choice of k. We characterize this property in the following result. For each S-valued random element Y, the average number of components of Y is naturally given by $\mathcal{K}(Y) = \sum_{k=1}^{\infty} k P(Y \in S_k)$.

Proposition 2. *Suppose $k > 1$ and $1 \leq p < \infty$. Let X be a H-valued random element, $\xi_j = \langle X, \phi_j \rangle$ and $X_k = \sum_{j=1}^{k} \xi_j \phi_j$. If $\{E(\|X - X_k\|_H^p)\}^{1/p} < \varepsilon$, then there exists an S-valued random element Y such that $\{E(\|X - Y\|_H^p)\}^{1/p} < \varepsilon$ and $\mathcal{K}(Y) < \mathcal{K}(X_k)$, provided that the probability density f_k of ξ_j is continuous at 0 and $f_k(0) > 0$.*

We note that the above result can be extended to the case $p = \infty$, where $\{E(\|Z\|_H^p)\}^{1/p}$ is replaced by $\inf\{w \in \mathbb{R} : P(\omega \in \Omega : \|Z(\omega)\|_H \leq w) = 1\}$.

2.2 Probability Densities on Mixture Inner Product Spaces

For S-valued random elements X, defining $K = K(X) = \sum_{k=1}^{\infty} k 1_{X \in S_k}$ and $X_k = \sum_{j=1}^{k} \langle X, \phi_j \rangle \phi_j$, then $X = \sum_{k=1}^{\infty} X_k 1_{K=k}$, and $X = X_k$ with probability $\pi_k = P(K = k)$. Since each X_k is of finite dimension, if the conditional density $f(X_k \mid K = k)$ exists for each k, then it is possible to define a probability density for X with respect to a base measure whose restriction to each S_k coincides with the k-dimensional Lebesgue measure. In contrast, for general random processes, it is well known that the small ball probability density does not exist (Li and Linde, 1999; Delaigle and Hall, 2010). An intuitive explanation is that with the mixture representation the probability mass of X is essentially concentrated on the mixture components S_k, each of which has a finite dimension, with high concentration on the leading components. The decay of the mixture proportions π_k as k increases then prevents the overall probability mass from escaping to infinity. Below we provide the details of this concept of a mixture density associated with MIPS.

It is well known that each H_k is isomorphic to $\mathbb{R}^k$, with associated Lebesgue measure τ_k. Defining a base measure $\tau(A) = \sum_{k=1}^{\infty} \tau_k(A \cap S_k)$ for $A \in \mathscr{B}(S)$, where we note that τ depends on the choice of the CONS, as change in the CONS leads to a different MIPS, the restriction of τ to each S_k is τ_k. Therefore, although τ itself is not a Lebesgue measure, the restriction to each finite-dimensional subspace H_k is.

For the random variables $\xi_j = \langle X, \phi_j \rangle$, $j \geq 1$, for a S-valued random element X assume that the conditional densities $f_k(\xi_1, \xi_2, \ldots, \xi_k) = f(\xi_1, \xi_2, \ldots, \xi_k \mid K = k)$ exist. With $\pi_k = P(X \in S_k) = P(K = k)$ we then define the mixture density function

$$f(x) = \sum_{k=1}^{\infty} \pi_k f_k(\langle x, \phi_1 \rangle, \langle x, \phi_2 \rangle, \ldots, \langle x, \phi_k \rangle) 1_{x \in S_k}, \quad \forall x \in S. \tag{1}$$

Note that even though there are infinitely many terms in (1), for any given realization $x = X(\cdot, \omega)$, only one of these terms is non-zero due to the presence of the indicator $1_{x \in S_k}$ and the fact that $X \in S$. Therefore, f is well defined for all $x \in S$ given $\sum_k \pi_k = 1$.

The presence of the indicator function $1_{x \in S_k}$ implies that the mixture density in (1) is distinct from any classical finite mixture model, where each component might have the same full support, while here the support of the each mixture component is specific to the component. The key difference to usual mixture models is that our model entails a mixture of densities that are defined on disjoint subsets, rather than on a common support. The following result implies that the problem of non-existence of a probability density in L^2 can be addressed by viewing functional data as elements of a mixture inner product space.

Theorem 2. *The measure τ is a σ-finite measure on S. In addition, if the conditional density $f_k(\xi_1, \xi_2, \ldots, \xi_k)$ exists for each k, then the probability distribution P_X on S induced by X is absolutely continuous with respect to τ. Moreover, the function f defined in* (1) *is a probability density of P_X with respect to τ.*

We note that the domain of f is S. Although S is dense in H, since f is not continuous, there is no natural extension of f to the whole space H. Nevertheless, we can extend both τ and f to H in the following straightforward way. Define the extended measure τ^* on H by $\tau^*(A) = \tau(A \cap S)$ for all $A \in \mathscr{B}(H)$. To extend f, we simply define $f(x) = 0$ if $x \in H \backslash S$. One can easily verify that τ^* is a measure on H extending τ, and f is a density function of X with respect to τ^*.

2.3 Constructing Mixture Inner Product Space Valued Random Elements

In this section, we focus on an important class of MIPS-valued random elements. Let $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$ be a sequence of uncorrelated centered random variables such that joint probability densities $\tilde{f}_k$ of $\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k$ exist for all k. Suppose K is a positive random integer with distribution $\pi = (\pi_1, \pi_2, \ldots)$ where K is independent of $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$, and $\pi_k = \Pr(K = k)$. Then we construct a random element $X = \mu + \sum_{k=1}^{K} \tilde{\xi}_k \phi_k$, where $\mu \in H$. We refer to a MIPS with random elements constructed in this way as a *generative MIPS*. Note that the mean element μ is allowed to be in the space H. Therefore, the centered process $X - \mu$, which is the primary object that the MIPS framework targets, takes value in a MIPS. This feature enhances the practical applicability of the MIPS framework. A generative MIPS has particularly useful properties that we discuss next.

In order to define mean and covariance of X, we also need that $E(\|X\|_H^2) < \infty$; a simple condition that implies this assumption is $\sum_{j=1}^{\infty} (\sum_{k=j}^{\infty} \pi_k) \mathrm{var}(\tilde{\xi}_j) < \infty$. Indeed, with $\pi_j^* = \sum_{k=j}^{\infty} \pi_k$,

$$\begin{aligned} E(\|X - \mu\|_H^2) &= E\left(\sum_{j=1}^{K} \tilde{\xi}_j^2\right) = EE\left(\sum_{j=1}^{K} \tilde{\xi}_j^2 \mid K\right) = \sum_{k=1}^{\infty} \pi_k E\left(\sum_{j=1}^{k} \tilde{\xi}_j^2\right) \\ &= \sum_{k=1}^{\infty} \pi_k \sum_{j=1}^{k} \mathrm{var}(\tilde{\xi}_j) = \sum_{j=1}^{\infty} \left(\sum_{k=j}^{\infty} \pi_k\right) \mathrm{var}(\tilde{\xi}_j) = \sum_{j=1}^{\infty} \pi_j^* \mathrm{var}(\tilde{\xi}_j) < \infty, \end{aligned}$$

$E(\|X\|_H^2) \le E(\|X - \mu\|_H^2) + \|\mu\|_H^2 < \infty$, and $(X - \mu)$ is seen to be a S-valued random element. Under the condition $E(\|X - \mu\|_H^2) < \infty$, $E(X - \mu) = 0$ and hence $E(X) = \mu$. Without loss of generality, we assume $\mu = 0$ in the following.

To analyze the covariance structure of $X = \sum_{k=1}^{K} \tilde{\xi}_k \phi_k$, consider $\xi_k = \langle X, \phi_k \rangle$. Then $\xi_k = \tilde{\xi}_k 1_{K \ge k}$, $E(\xi_k) = 0$, $\mathrm{var}(\xi_k) = \pi_k^* \mathrm{var}(\tilde{\xi}_k)$, and $E(\xi_j \xi_k) = 0$, and $\xi_1, \xi_2, \ldots$ are seen to be uncorrelated centered random variables with variance $\pi_k^* \mathrm{var}(\tilde{\xi}_k)$. Furthermore, because K is independent of the $\tilde{\xi}_k$, the conditional density of $\xi_1, \xi_2, \ldots, \xi_k$ given $K = k$ is the joint density of $\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k$. If $E(\|X\|_H^2) < \infty$, the covariance operator Γ for X exists (Hsing and Eubank, 2015). The ϕ_k are eigen-elements of Γ, as

$$\begin{aligned} \Gamma \phi_k &= E(X \langle X, \phi_k \rangle) = E(X \xi_k) = E(X \tilde{\xi}_k 1_{K \ge k}) = EE(X \tilde{\xi}_k 1_{K \ge k} \mid K) \\ &= \sum_{j=1}^{\infty} \pi_j E(X \tilde{\xi}_k 1_{K \ge k} \mid K = j) = \sum_{j=k}^{\infty} \pi_j E(\tilde{\xi}_k \sum_{m=1}^{j} \tilde{\xi}_m \phi_m) = \pi_k^* \mathrm{var}(\tilde{\xi}_k) \phi_k, \end{aligned} \tag{2}$$

where the last equality is due to uncorrelatedness of $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$. From (2), the eigenvalue λ_k corresponding to the k-th eigen-element ϕ_k is

$$\lambda_k = \pi_k^* \mathrm{var}(\tilde{\xi}_k). \tag{3}$$

Since $\phi_1, \phi_2, \ldots$ is a CONS of H, Γ has no other eigen-element in H. Therefore, Γ admits the eigen decomposition $\Gamma = \sum_{k=1}^{\infty} \lambda_k \phi_k \otimes \phi_k$, where $(x \otimes y) z = \langle x, z \rangle y$ for $x, y, z \in H$. For the special case where $H = L^2$, this feature establishes a connection to traditional functional principal component analysis and suggests implementation of MIPS in this special case by adopting well studied functional principal component analysis methods; see the next section for details.

An important consequence of these considerations is that for each random element $\tilde{X} \in H$ with $E(\|\tilde{X}\|_H^2) < \infty$ and for which the covariance operator $\tilde{\Gamma}$ has an eigendecomposition $\tilde{\Gamma} = \sum_{k=1}^{\infty} \tilde{\lambda}_k \tilde{\phi}_k \otimes \tilde{\phi}_k$ (assuming w.l.o.g. that $\tilde{\phi}_1, \tilde{\phi}_2, \ldots$ form a CONS of H), there exists a MIPS $\tilde{S}$ and a $\tilde{S}$-valued random element Z, such that the covariance operator Γ of Z has the same set of eigen-elements. To see this, define $\tilde{S}$ to be the MIPS generated by $\tilde{\phi}_1, \tilde{\phi}_2, \ldots$ and note that $\zeta_k = \langle \tilde{X}, \tilde{\phi}_k \rangle, k \geq 1$, are uncorrelated random variables with variances $\tilde{\lambda}_k$ (Hsing and Eubank, 2015). Choose an independent random positive integer K with distribution $\pi = (\pi_1, \pi_2, \ldots)$ and $\pi_k > 0$ for all k, and set $Z = \sum_{k=1}^{K} \zeta_k \tilde{\phi}_k$. Since $\sum_{j=1}^{\infty} \pi_j^* \mathrm{var}(\zeta_j) \leq \sum_{j=1}^{\infty} \mathrm{var}(\zeta_j) < \infty$, we have $E(\|Z\|_H^2) < \infty$. Therefore, the derivation in (2) applies to Z and shows that the covariance operator of Z has exactly the eigen-elements $\tilde{\phi}_1, \tilde{\phi}_2, \ldots$.

3 Application to Functional Data Analysis

In this section we demonstrate how the MIPS concept can be applied to provide a new approach for the analysis of functional data and demonstrate how this can be utilized to define a density function for functional data via the density function (1), thus providing a new perspective on one of the foundational issues of functional data analysis. In the following sections, we show how to apply the theory of MIPS to analyze functional data.

3.1 L^p-Denseness of Mixture Inner Product Space Valued Random Processes

Theorem 1 implies that any given H-valued random element can be uniformly approximated by MIPS-valued random elements. An important consequence is that the set of MIPS-valued random elements is dense in an L^p sense, as follows. For $1 \leq p < \infty$, let $L^p(\Omega, \mathscr{E}, P; H)$ be the space of H-valued random elements X such that $E(\|X\|_H^p) < \infty$. It is well known (Vakhania *et al.*, 1987) that $L^p(\Omega, \mathscr{E}, P; H)$ (with elements defined as equivalence classes) is a Banach space with norm $\|X\|_{L^p} = \{E(\|X\|_H^p)\}^{1/p}$ for every $X \in L^p(\Omega, \mathscr{E}, P; H)$, where for $p = \infty$, $L^\infty(\Omega, \mathscr{E}, P; H)$ denotes the Banach space with the essential supremum norm. Since each S-valued random element is also an H-valued random element according to Proposition 1, the space $L^p(\Omega, \mathscr{E}, P; S)$ is a subspace of $L^p(\Omega, \mathscr{E}, P; H)$. The following corollary states that $L^p(\Omega, \mathscr{E}, P; S)$ is dense in $L^p(\Omega, \mathscr{E}, P; H)$, which is an immediate consequence of Theorem 1.

Corollary 1. *If X is a H-valued random element and S is a MIPS of H, there exists a sequence of S-valued random elements $Y_1, Y_2, \ldots$, such that $\|X - Y_j\|_{L^p} \to 0$ as $j \to \infty$ for $1 \leq p \leq \infty$, i.e., $L^p(\Omega, \mathscr{E}, P; S)$ is a dense subset of $L^p(\Omega, \mathscr{E}, P; H)$.*

Applying this result to the Hilbert space $H = L^2(D)$, which is the set of real functions $f : D \to \mathbb{R}$ such that $\int_D |f(t)|^2 dt < \infty$, where D is a compact subset of $\mathbb{R}$, e.g. $D = [0, 1]$, we conclude that the set of MIPS-valued random processes is dense in the space of all $L^2(D)$ random processes. This denseness implies that when modelling functional data with MIPS-valued random processes, the results are arbitrarily close to those one would have obtained with the traditional L^2 based functional data analysis approaches in the L^2 sense. A major difference between the two approaches is that each functional element is always finite-dimensional in the MIPS framework, as it belongs to one of the subspaces S_k, where the MIPS is $S = \bigcup_{k=1}^{\infty} S_k$, as defined above, while in the classical L^2 framework each element is infinite-dimensional. The denseness of MIPS in L^2 provides additional justification for the adoption of this new approach.

As we will demonstrate below, modeling functional data in the MIPS framework enjoys extra flexibility and parsimony in representing observed functional data. And, as mentioned before, it provides a way to define probability densities for functional data within the full MIPS space, avoiding ad hoc truncation approaches to which one must resort when tackling the density problem directly in the traditional functional data space L^2.

3.2 Model and Estimation

In the following, we develop a MIPS based functional mixture model from a practical modeling perspective. A practical motivation to adopt a mixture model is that it enables adaptive choice of the number of components that are included to represent a given functional trajectory. This adaption is with respect to the complexity of the trajectory that is to be represented. The basic idea is that trajectories that have more features and shape variation relatively to other trajectories require a larger number of components to achieve a good representation, while those that are flat and have little shape variation will require fewer components. This contrasts with the "one size fits all" approach of functional principal component analysis or other expansions in basis functions, where the expansion series always includes infinitely many terms.

To implement MIPS based shape-adaptive modeling, we require that the process X underlying the observed data is a generative MIPS random process, as defined in the beginning of Section 2.3. This requirement is formalized in the following assumption.

(A0) The observed functional data $X_1, X_2, \ldots, X_n$ are i.i.d. realizations of a generative MIPS random process X, i.e., $X = \sum_{k=1}^{K} \tilde{\xi}_k \phi_k$ for a sequence of uncorrelated centered random variables $\tilde{\xi}_k$ and a CONS $\Phi = \{\phi_1, \phi_2, \ldots\}$ of the space L^2. Here K is an integer valued random variable that is independent of the $\tilde{\xi}_k$ and has the distribution $P(K = k) = \pi_k > 0$ for a sequence π_k, $k \geq 1$, such that $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$. The sequence $\lambda_k = \mathrm{var}(\tilde{\xi}_k)$ is strictly decreasing with increasing k, i.e., $\lambda_1 > \lambda_2 > \cdots$.

Under this assumption, the CONS $\Phi = \{\phi_1, \phi_2, \ldots\}$ is exactly the one featured in the Karhunen-Loève decomposition, according to the discussion in Section 2.3.

We demonstrate that the probability density for functional mixtures as defined in the previous section can be consistently estimated under suitable regularity conditions. For convenience, the functions $X_1,\ldots,X_n$ are considered fully observed; densely sampled functional data can be individually pre-smoothed to produce asymptotically equivalent estimates (Hall *et al.*, 2006). The eigenfunctions $\phi_1,\phi_2,\ldots$ and associated eigenvalues are unknown and need to be estimated from the data.

We consider the case where the component densities f_k, $k \geq 1$, are parametrizable by set of parameters $\theta_k = (\theta_{k,1},\theta_{k,2},\ldots,\theta_{k,d_k})^T \in \mathbb{R}^{d_k}$ for some sequence of integers d_k, and the distribution of the mixture proportions π is governed by a sequence of parameters $\theta_\pi \in \mathbb{R}^{d_\pi}$. Although we develop here a parametric approach to model each component density f_k, it is worth noting that a nonparametric approach might be taken in a similar spirit as the nonparametric estimation of finite mixtures, developed by Benaglia *et al.* (2009) and Levine *et al.* (2011), where substantial additional technicalities are involved. We leave this topic for future research.

For notational convenience, let $\theta_{[k]}$ be the sequence of parameter sets, containing distinct parameters $\theta_1,\theta_2,\ldots,\theta_k$, as well as the parameters in θ_π that determine $\pi_1,\ldots,\pi_{k-1}$, with the dimension denoted by $d_{[k]}$, where $\theta_{[\infty]}$ stands for the sequence of all distinct parameters, such that

$$\theta_{[k]} = (\theta_{[\infty],1},\theta_{[\infty],2},\ldots,\theta_{[\infty],d_{[k]}}), \tag{4}$$

where $\theta_{[\infty],j}$ is the jth coordinate of $\theta_{[\infty]}$. Let $I_{[\infty],j}$ denote the domain of $\theta_{[\infty],j}$, and define $\Theta = \prod_{j=1}^{\infty} I_{[\infty],j}$. We write $f_k(\cdot)$ as $f_k(\cdot \mid \theta_k)$ to emphasize the dependence on θ_k. Similarly, $f(\cdot)$ is written as $f(\cdot \mid \theta_{[\infty]})$ to indicate that f is parametrized by $\theta_{[\infty]}$, i.e.,

$$f(X(\omega) \mid \theta_{[\infty]}) = \sum_{k=1}^{\infty} \pi_k f_k(\xi_1(\omega),\xi_2(\omega),\ldots,\xi_k(\omega) \mid \theta_k) 1_{X(\omega)\in S_k}. \tag{5}$$

For a generic parameter θ, we use θ_0 to denote its true value, and $\hat{\theta}$ to denote corresponding maximum likelihood estimators, e.g., $\theta_{[\infty],0}$ denotes the true parameters of $\theta_{[\infty]}$.

To illustrate the key idea, we make the simplifying assumption of compactness of the parameter space, which may be relaxed by introducing more technicalities. The condition below characterizes the compactness of the parameter space $\Theta = \prod_{j=1}^{\infty} I_{[\infty],j}$ as a product of compact spaces, using Tychonoff's theorem.

(A1) For each $j = 1,2,\ldots$, $I_{[\infty],j}$ is a non-empty compact subset of $\mathbb{R}$, and thus $\Theta = \prod_{j=1}^{\infty} I_{[\infty],j}$ is compact (by Tychonoff's theorem).

With eigenfunctions $\phi_1,\phi_2,\ldots$ estimated by decomposing the sample covariance operator, the principal component scores ξ_{ik} are estimated by $\hat{\xi}_{ik} = \langle X_i, \hat{\phi}_k\rangle$ for each $i = 1,2,\ldots,n$ and $k = 1,2,\ldots$, where $\hat{\phi}_k$ are the standard estimates of ϕ_k. To quantify the estimation quality, we postulate a standard regularity condition for X (Hall and Hosseini-Nasab, 2006) and a polynomial decay assumption for the eigenvalues $\lambda_1 > \lambda_2 > \cdots > 0$ (Hall and Horowitz, 2007).

(A2) For all $C > c'$ and some $\varepsilon' > 0$, where $c' > 0$ is a constant, $\sup_{t\in D} E\{|X(t)|^C\} < \infty$ and $\sup_{s,t\in D} E[\{|s-t|^{-\varepsilon'}|X(s)-X(t)|\}^C] < \infty$.

(A3) For all $k \geq 1$, $\lambda_k - \lambda_{k+1} \geq C_0 k^{-b-1}$ for constants $C_0 > 0$ and $b > 1$, and also $\pi_k = O(k^{-\beta})$ for a constant $\beta > 1$.

Note that $\sum_k \lambda_k < \infty$ and $\sum \pi_k = 1$ imply $b > 1$ and $\beta > 1$ and one also has $\pi_k^* = \sum_{j=k}^{\infty} \pi_k = O(k^{-\beta+1})$. Condition (A3) also implies that $\lambda_k \geq C' k^{-b}$ for a constant $C' > 0$ for all k. Therefore, if $\rho_k = \text{var}(\tilde{\xi}_k)$, the relation $\lambda_k = \pi_k^* \text{var}(\tilde{\xi}_k)$ that was derived in (3) implies $\rho_k = \lambda_k / \pi_k^* \geq C_\rho k^{-b+\beta-1}$ for a constant $C_\rho > 0$ and for all k. Note that the case $-b+\beta-1 > 0$, for which the variances of the $\tilde{\xi}_k$ diverge, is not excluded.

Our next assumption concerns the regularity of mixture components $f_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)$ and $g_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k) = \log f_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)$, where the dependence on θ_k is suppressed when no confusion arises.

(A4) For $k = 1, 2, \ldots$, $f_k(\cdot \mid \theta_k)$ is continuous at all arguments θ_k. There exist constants $C_1, C_2, C_3 \geq 0$, $-\infty < \alpha_1, \alpha_2 < \infty$, $0 < \nu_1 \leq 1$, $0 < \nu_2 \leq 2$ and functions $H_k(\cdot)$ such that, for all $k = 1, 2, \ldots,$ g_k satisfies $|g_k(u) - g_k(v)| \leq C_1 H_k(v) \|u - v\|^{\nu_1} + C_2 k^{\alpha_2} \|u - v\|^{\nu_2}$ for all $u, v \in \mathbb{R}^k$, and $E\{H_k(\xi_1, \xi_2, \ldots, \xi_k)\}^2 \leq C_3 k^{2\alpha_1}$.

In the following, we use $\alpha = \max\{\alpha_1, \alpha_2\}$ and $\nu = \min(2\nu_1, \nu_2)$. Note that Hölder continuity is a special case for $C_1 = 0$. Given (A3), one can verify that (A4) is satisfied for the case of Gaussian component densities with $C_1, C_2, C_3 > 0$, $\nu_1 = 1$, $\nu_2 = 2$, $\alpha_1 > 2^{-1} \max(1-b, 2b-3\beta+4)$ and $\alpha_2 = \max(0, b-\beta+1)$. The condition on $|g_k(u) - g_k(v)|$ in (A4) implicitly assumes a certain growth rate of $d_{[k]}$ as k goes to infinity. For instance, $E\{H_k(\xi_1, \xi_2, \ldots, \xi_k)\}^2$ is a function of the parameter set $\theta_{[k]}$. By the compactness assumption on $\theta_{[\infty]}$, the parameters have a common upper bound. With this upper bound, $E\{H_k(\xi_1, \xi_2, \ldots, \xi_k)\}^2$ can be bounded by some function R of $d_{[k]}$. By postulating $E\{H_k(\xi_1, \xi_2, \ldots, \xi_k)\}^2 \leq C_3 k^{2\alpha_1}$ in (A4), we implicitly assert that the function $R(d_{[k]})$ can be bounded by a polynomial of k, with the exponent $2\alpha_1$. We would need a larger value for α_1 when $d_{[k]}$ grows faster with k. A similar argument applies to α_2.

To state the needed regularity conditions for the likelihood function, we need some notations. Let $Q_r = \min(K, r)$, and define $Z_i = \sum_{j=1}^{Q_r} \langle X_i, \phi_j \rangle \phi_j$, so that $Z_i \in S_q$, $q \leq r$. The log-likelihood of a single observation Z is

$$L_{r,1}(Z \mid \theta_{[r]}) = \log \left\{ (1 - \sum_{k=1}^{r-1} \pi_k) f_r(Z \mid \theta_{[r]}) 1_{Z \in S_r} + \sum_{k=1}^{r-1} \pi_k f_k(Z \mid \theta_{[k]}) 1_{Z \in S_k} \right\}. \tag{6}$$

The log-likelihood function of $\theta_{[r]}$ for a sample $Z_1, \ldots, Z_n$ accordingly is

$$L_{r,n}(\theta_{[r]}) = n^{-1} \sum_{i=1}^{n} L_{r,1}(Z_i \mid \theta_{[r]}), \tag{7}$$

with maximizer $\hat{\theta}_{[r]}$. We impose the following regularity condition on $L_r(\theta_{[r]}) = E\{L_{r,1}(Z \mid \theta_{[r]})\}$.

(A5) There exist constants $h_1, h_2, h_3, a_1, a_2, a_3 > 0$ such that for all $r \geq 1$, $U_r = \{\theta_{[r]} : L_r(\theta_{[r],0}) - L_r(\theta_{[r]}) < h_1 r^{-a_1}\}$ is contained in a neighborhood $\mathcal{B}_r = \{\theta_{[r]} : \|\theta_{[r],0} - \theta_{[r]}\| < h_2 r^{-a_2}\}$ of $\theta_{[r],0}$, where $\theta_{[r],0}$ denotes the true parameters of $\theta_{[r]}$. Moreover, $L_r(\theta_{[r],0}) - L_r(\theta_{[r]}) \geq h_3 r^{-a_3} \|\theta_{[r],0} - \theta_{[r]}\|^2$ for all $\theta_{[r]} \in U_r$.

Writing $a = \max\{a_1, a_3\}$, we observe that (A5) is satisfied when each component f_k is Gaussian for any $a > 1$, and (A1) and (A3) hold. (A5) essentially states that the global maximizer of L_r is unique and uniformly isolated from other local maximizers with an order r^{-a}. Such a condition on separability is necessary when there are infinitely many mixture components in a model. We note that (A5) also ensures identifiability of the global maximizer.

The next assumption is used to regulate the relationship between the mixture proportions π_k and the magnitude of $g_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)$, by imposing a bound on g_k for increasing k.

(A6) For a constant $c < \beta - 1$, $E|g_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)| = O(k^{c-a})$, where a is defined in (A5) and β in (A3) and $g_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k) = \log f_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)$.

The constraint $\beta > c + 1$ in (A6) guarantees that in light of $\pi_k = O(k^{-\beta})$, as per (A3), the mixture proportions π_k decay fast enough relative to average magnitude of $g_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)$ to avoid a singularity that might arise in the summing operation to construct the density f in (1) when the magnitude of $g_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)$ grows too fast. This bound will prevent that too much mass is allocated to the components with higher dimensions in the composite mixture density f. Such a scenario would preclude the existence of a density and is avoided by tying the growth of g_k to the decline rate of the π_k, as per (A6).

From a practical perspective, faster decay of the π_k that places more probability mass on the lower-order mixture components will help stabilize the estimation procedure, as it is difficult to estimate the high-order eigenfunctions that are needed for the higher order components. For the case of Gaussian component densities, a simple calculation gives $E|g_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)| = O(k \log k)$, thus (A6) is fulfilled for any $c > a + 1$. This will also imply that $\beta > a + 2$. An extreme situation arises when $\pi_k = 0$ for $k \geq k_0$ for some $k_0 > 0$, i.e., the dimension of the functional space is finite and the functional model essentially becomes parametric. In this case the construction of the mixture density in functional space is particularly straightforward.

The following theorem establishes estimation consistency for a growing sequence of parameters $\theta_{[r_n]}$ as the sample size n increases, and consequently the consistency of the estimated probability density at any functional observation $x \in S$ as argument. Define constants $\gamma_1 = (2b+3)\nu/2 + \alpha - 2\beta$, that $\gamma_2 = a + (\gamma_1 + 2)1_{\{\gamma_1 > -2\}}$, and set $\gamma = \min\{\nu/(2\gamma_2), 1/(2b+2)\}$.

Theorem 3. *If assumptions (A0)-(A6) hold and $r_n = O(n^{\gamma - \varepsilon})$ for any $0 < \varepsilon \leq \gamma$, then the global maximizer $\hat{\theta}_{[r_n]}$ of $\hat{L}_{r,n}(\theta_{[r_n]})$ satisfies*

$$\|\hat{\theta}_{[r_n]} - \theta_{[r_n],0}\| \xrightarrow{p} 0,$$

where $\theta_{[r_n]}$ *is defined in* (4)*, and* $\hat{L}_{r,n}(\theta_{[k]})$ *is the likelihood function obtained by plugging the estimated quantities* $\hat{\phi}_k$ *and* $\hat{\xi}_{ik}$ *into* $L_{r,n}(\theta_{[r_n]})$ *defined in* (7)*. Consequently, for any* $x \in S = \bigcup_{k=1}^{\infty} S_k$*, one has*

$$\left| f(x \mid \hat{\theta}_{[r_n]}) - f(x \mid \theta_{[\infty],0}) \right| \xrightarrow{p} 0,$$

where f *is the mixture probability density defined in* (5)*.*

We see from Theorem 3 that the number of consistently estimable mixture components, r_n, grows with a polynomial rate in terms of the sample size n. From (A4), the proximity to a singularity of the component density f_k is seen to increase as α increases, indicating more difficulty in estimating f_k, and thus restricting the rate of increase in r_n. Faster decay rates of the eigenvalues λ_k, relative to the decline rates in the mixture proportions π_k and quantified by b and $(b-\beta+1)$ respectively, lead to limitations in the number of eigenfunctions that can be reliably estimated and this is reflected in a corresponding slowing of the rise of the number of mixture components r_n that can be included. The rate at which r_n can increase also depends on the decay of the π_k as quantified by β.

3.3 Fitting Algorithm

We present an estimation method based on the expectation-maximization algorithm to determine the mixture probabilities π_k for $k = 1,2\ldots$, and the number K_i of components that are associated with each individual trajectory X_i. For simplicity, we assume that the mixture proportions π_k are derived from a known family of discrete distributions that can be parametrized with one or few unknown parameters, denoted here by ϑ, simplifying the notation introduced in Section 3.2. A likelihood for fitting individual trajectories with K components can then be constructed. The following algorithm is based on fully observed X_i. Modifications for the case of discretely observed data are discussed at the end of the section.

To be specific, we outline the algorithm for the mixture density of Gaussian processes, and use $\pi \sim \text{Poisson}(\vartheta)$, i.e., $P(K = k \mid \vartheta) = \vartheta^k e^{-\vartheta}/k!$. Versions for other distributions can be developed analogously. Assume that $X_1,\ldots,X_n$ are centered without loss of generality. Projecting X_i onto each eigenfunction ϕ_j, we obtain the functional principal component scores $\xi_{i1},\xi_{i2},\ldots$ of X_i. Given $K_i = k$, $(\xi_{i1},\xi_{i2},\ldots,\xi_{ik}) = (\tilde{\xi}_{i1},\ldots,\tilde{\xi}_{ik})$ and

$$(\tilde{\xi}_{i1},\ldots,\tilde{\xi}_{ik})^T \sim N\big(0,\Sigma_\rho^{(k)}\big), \tag{8}$$

where $\Sigma_\rho^{(k)}$ is the $k\times k$ diagonal matrix with diagonal elements $\rho_j = \text{var}(\tilde{\xi}_j)$, $j = 1,2,\ldots,k$. The likelihood $f(X_i \mid K_i = k)$ of X_i conditional on $K_i = k$ is then given by

$$f(X_i \mid K_i = k) = \frac{1}{\sqrt{(2\pi)^k|\Sigma_\rho^{(k)}|}} \exp\left[-\frac{1}{2}(\xi_{i1},\ldots,\xi_{ik})(\Sigma_\rho^{(k)})^{-1}(\xi_{i1},\ldots,\xi_{ik})^T\right]. \tag{9}$$

Note that one needs the eigenvalues ρ_k to characterize the distribution of the observations X_i given

K_i. Based on equation (3), one can adopt standard functional principal component analysis for the entire sample that contains realizations of X, i.e., extract the eigenvalues λ_k of G first and then utilize $\rho_k = \lambda_k/\pi_k^*$. This however requires to infer the unknown mixture proportions π_k. To address this conundrum, we treat K_i as a latent variable or missing value and adopt the expectation-maximization paradigm, as follows.

1. Obtain consistent estimates $\hat{\phi}_k(\cdot)$ of $\phi_k(\cdot)$ and $\hat{\lambda}_k$ of λ_k, $k = 1, 2, \ldots$, from functional principal component analysis by pooling the data from all individuals, following well-known procedures (Dauxois *et al.*, 1982; Hall and Hosseini-Nasab, 2006), followed by projecting each observation X_i onto each $\hat{\phi}_k$ to obtain estimated functional principal component scores $\hat{\xi}_{ik}$. As starting value for the Poisson parameter ϑ, we set $\vartheta = k$, where k is the smallest integer such that the fraction of variation explained by the first k principal components exceeds 95%.

2. Plug in the estimate $\hat{\lambda}_k$ for λ_k and calculate $\hat{\rho}_k = \hat{\lambda}_k/\pi_k^*$, with $\pi_k = p(k \mid \vartheta)$ based on the current estimate of ϑ, which we denote by $\vartheta^{(t)}$. Obtain the conditional expectation of K_i given X_i,

$$E(K_i \mid X_i) = \frac{\sum_{k=1}^{\infty} k f(X_i \mid K_i = k) P(K_i = k \mid \vartheta^{(t)})}{\sum_{k=1}^{\infty} f(X_i \mid K_i = k) P(K_i = k \mid \vartheta^{(t)})}, \tag{10}$$

   where $f(X_i \mid K_i = k)$ is given by (9). It is natural to use the nearest integer, denoted by $E_i(K_i \mid X_i)$. The updated estimate of ϑ is given by $\vartheta^{(t+1)} = n^{-1} \sum_{i=1}^{n} E_i(K_i \mid X_i)$. Repeat this step until $\vartheta^{(t)}$ converges. By the ascent property of the EM algorithm, $\vartheta^{(t)}$ converges to a local maximizer. In practice, this step is repeated until a specified convergence threshold is reached that may be defined in terms of the relative change of ϑ, i.e., $|\vartheta^{(t+1)} - \vartheta^{(t)}|/\vartheta^{(t)}$.

3. Each X_i is represented by $X_i = \sum_{j=1}^{K_i} \hat{\xi}_{ij} \hat{\phi}_j$, where K_i is obtained as in (10).

In the numerical implementation it is advantageous to only keep the positive eigenvalue estimates $\hat{\rho}_k^+$, and to introduce a truncated Poisson distribution that is bounded by $K_n^+ = \max\{k : \hat{\rho}_k^+ > 0\}$,

$$p^+(k \mid \vartheta, K_n^+) = \frac{\vartheta^k}{k!(\sum_{\ell=0}^{K_n^+} \vartheta^\ell/\ell!)} \equiv \pi_k^+, \qquad k = 0, 1, \ldots, K_n^+. \tag{11}$$

Since the maximum likelihood estimate of ϑ in (11) based on the truncated Poisson distribution is complicated and does not have an analytical form, it is expedient to numerically maximize the conditional expectation of the log-likelihood with respect to ϑ given the observed data X_i, $i = 1, \ldots, n$, and the current estimate $\vartheta^{(t)}$,

$$\begin{aligned} & \sum_{i=1}^{n} E\{\log p^+(K_i \mid \vartheta, K_n^+) \mid X_i, \vartheta^{(t)}\} \\ = \quad & \sum_{i=1}^{n} \frac{\sum_{k=1}^{K_n^+} \log p^+(k \mid \vartheta, K_n^+) f(X_i \mid K_i = k) p^+(k \mid \vartheta^{(t)}, K_n^+)}{\sum_{k=1}^{K_n^+} f(X_i \mid K_i = k) p^+(k \mid \vartheta^{(t)}, K_n^+)}, \end{aligned} \tag{12}$$

and to consider the modified eigenvalues $\rho_k^+ = \lambda_k^+ / (\sum_{j=k}^{K_n^+} \pi_j^+)$.

In many practical situations the trajectories X_i are measured at a set of discrete points $t_{i1}, \ldots, t_{im_i}$, rather than fully observed. This situation requires some modifications of the estimation procedures. For step 1, the eigenfunctions ϕ_k, $k = 1, 2, \ldots$, can be consistently estimated via a suitable implementation of functional principal component analysis, where for this estimation step unified frameworks have been developed for densely or sparsely observed functional data (Li and Hsing, 2010; Zhang and Wang, 2016). If the design points are sufficiently dense, alternatively, individual smoothing as a preprocessing step may be applied and one may then treat the pre-smoothed functions $\hat{X}_1, \ldots, \hat{X}_n$ as if they were fully observed.

In situations where the measurements are noisy, a possible approach is to compute the likelihoods conditional on the available observations $U_i = (U_{i1}, \ldots, U_{im_i})$, where $U_{ij} = X_i(t_{ij}) + \varepsilon_{ij}$ with measurement errors ε_{ij} that are independently and identically distributed according $N(0, \sigma^2)$ and independent of X_i. Under joint Gaussian assumptions on $X_i^{(k)}$ and the measurement errors, the $m_i \times m_i$ covariance matrix of U_i is

$$\operatorname{cov}(U_i \mid k) = \left\{ \sum_{r=1}^{k} \rho_r \phi_r(t_{ij}) \phi_r(t_{i\ell}) \right\}_{1 \le j, \ell \le m_i} + \sigma^2 I_{m_i} \equiv \Sigma_{U_i}^{(k)}, \tag{13}$$

where I_{m_i} denotes the $m_i \times m_i$ identity matrix. The likelihood $f(U_i \mid K)$ is then derived from $N\left(\mu_i, \Sigma_{U_i}^{(k)}\right)$ with $\mu_i = \{\mu(t_{i1}), \ldots, \mu(t_{im_i})\}^\top$ and the estimation procedure is modified by replacing $f(X_i \mid K_i = k)$ with $f(U_i \mid K_i = k)$ in equation (10). The following modifications are applied at steps 1 and 3: in step 1, the projections of the X_i onto the ϕ_k are skipped; in step 3, the functional principal component scores ξ_{ik}, $k = 1, \ldots, K_i$, are obtained in a final step by numerical integration for the case of densely sampled data, $\xi_{ik} = \int X_i(t) \phi_k(t) dt$, plugging in eigenfunction estimates $\hat{\phi}_k$, or by PACE estimates for the case of sparse data (Yao *et al.*, 2005).

4 Simulation Study

To demonstrate the performance of the proposed mixture approach, we conducted simulations for four different settings. For all settings, the simulations are based on $n = 200$ trajectories from an underlying process X with mean function $\mu(t) = t + \sin(t)$ and covariance function derived from the Fourier basis $\phi_{2\ell-1} = \cos\{(2\ell-1)\pi t/10\}/\sqrt{5}$ and $\phi_{2\ell} = \sin\{(2\ell-1)\pi t/10\}/\sqrt{5}$, $\ell = 1, 2, \ldots, t \in T = [0, 10]$. For $i = 1, \ldots, n$, the ith trajectory was generated as $X_i(t) = \mu(x) + \sum_{k=1}^{K_i} \xi_{ik} \phi_k(t)$. Two different cases for ξ_{ik} were considered. One is Gaussian, where $\xi_{ik} \sim N(0, \rho_k)$ with $\rho_k = 16k^{-1.8}$. The other is non-Gaussian, where ξ_{ik} follows a Laplace distribution with mean zero and variance $16k^{-1.8}$, which is included to illustrate the effect of mild deviations from the Gaussian case. Each trajectory was sampled at $m = 200$ equally spaced time points $t_{ij} \in T$, and measurements were contaminated with independent measurement errors $\varepsilon_{ik} \sim N(0, \sigma^2)$, i.e., the actual observations are $U_{ij} = X_i(t_{ij}) + \varepsilon_{ij}$,

$j = 1, \ldots, m$. Two different levels were considered for σ^2, namely, 0.1 and 0.25.

The four settings differ in the choice of the latent trajectory dimensions K_i. In the *multinomial* setting, K_i is independently sampled from a common distribution $(\pi_1, \ldots, \pi_{15})$, where the event probabilities $\pi_1, \ldots, \pi_{15}$ are randomly generated according to a Dirichlet distribution. In the *Poisson* setting, each K_i is independently sampled from a Poisson distribution with mean $\vartheta = 6$. In the *finite* setting, each K_i is set to a common constant equal to 12, and in the *infinite* setting, each K_i is set to a large common constant equal to 25, which mimics the infinite nature of the process X. In the multinomial and Poisson settings the K_i vary from subject to subject, while in the finite and infinite settings, they are the same across all subjects. In the multinomial and finite settings, the $K_1, \ldots, K_n$ are capped by a finite number that does not depend on n, whereas in the Poisson and infinite settings the K_i are in principle unbounded and can be arbitrarily large. In our implementation, we used the Gaussian-Poisson fitting algorithm described in Section 3.3 to obtain fits for the generated data in all four settings.

For evaluation purposes, we generated a test sample of size 20000 for each setting. The population model components, such as the mean, covariance, eigenvalues and eigenfunctions and also the rate parameter ϑ were estimated from the training sample, while the subject-level estimates, K_i and the estimates of the functional principal component were obtained from the generated data $\{U_{ij}^*, j = 1, \ldots, m\}$ that are observed for the i-th subject in the test set X_i^*. Of primary interest is to achieve good trajectory recovery with the most parsimonious functional data representation possible, using as few components as possible to represent each trajectory. The performance of the trajectory recovery is measured in terms of the average integrated squared error obtained for the trajectories in the test set, $\text{AISE} = n^{-1} \sum_{i=1}^n \int_T \{X_i^*(t) - \widehat{X}_i^*(t)\}^2 dt$. The parsimoniousness of the representations is quantified by the average number of principal components $K_{\text{avg}} = n^{-1} \sum_{i=1}^n K_i$ that are chosen for the subjects. For the traditional functional principal component analysis this is always a common choice of $K_i = K$ for all subjects. The results are presented in Table 1. For comparison, the minimized average integrated squared error for functional principal component analysis with its common choice K for the number of components across all trajectories is also included in the last column.

The results clearly show that in both Poisson and multinomial settings the proposed mixture method achieves often substantially smaller average integrated squared errors while utilizing fewer components on average than the traditional functional principal component analysis. In contrast, in the fixed and infinite settings, the proposed mixture method recovers trajectories with an error that is comparable to that of traditional functional principal component analysis, using roughly the same number of principal components. We conclude that the proposed mixture model is substantially better in some situations where trajectories are not homogeneous in terms of their structure, while the price to be paid for situations where the standard functional principal component analysis is the preferred approach is relatively small. We also note that a mild deviation from the Gaussian assumption does not have much impact on the performance. We also ran additional simulations for the *Poisson* settings with $\sigma^2 = 0.1$ and different sample sizes. In comparison to the true value $\vartheta = 6$ and the estimate $\hat{\vartheta} = 6.78(0.14)$ for $n = 200$, the estimates $6.39(0.13), 6.21(0.10), 6.12(0.06), 6.06(0.04)$ for

$n = 500, 1000, 2000, 5000$, respectively, provide empirical support for estimation consistency, where the standard errors in parentheses are based on 100 Monte Carlo runs.

5 Application

Longitudinal data on daily egg-laying for female medflies, Ceratitis Capitata, were obtained in a fertility study as described in Carey *et al.* (1998). The data set is available at `http://anson.ucdavis.edu/∼mueller/data/medfly1000.html`. Selecting flies that survived for at least 25 days to ensure that there is no drop-out bias yielded a subsample of $n = 750$ medflies. For each of the flies one has then trajectories corresponding to the number of daily eggs laid from birth to age 25 days. Shown in the top-left panel of Figure 1 are the daily egg-laying counts of 50 randomly selected flies. We apply a square-root transformation to the egg counts to symmetrize the errors as a pre-processing step. Applying standard functional principal component analysis yields estimates of the mean, covariance and eigenfunctions, as shown in the last three panels of Figure 1.

Visual inspection indicates that the egg-laying trajectories possess highly variable shapes with different varying numbers of local modes. This motivates us to apply the proposed functional mixture model. The goal is to parsimoniously recover the complex structure of the observed trajectories. For evaluation, we conduct 100 runs of 10-fold cross-validation, where in each run, we shuffle the data independently, and use 10% of the flies as validation set for obtaining the subject-level estimates, which include the latent dimensions K_i and the functional principal component scores, and use the remaining 90% of the flies as training set. The resulting cross-validated relative squared errors are

$$\text{CVRSE} = n^{-1} \sum_{l=1}^{10} \sum_{i \in D_l} \left([\sum_{j=1}^{m} \{U_{ij} - \widehat{X}_i^{-D_l}(t_{ij})\}^2] / \sum_{j=1}^{m} U_{ij}^2 \right),$$

where D_l is the lth validation set containing 10% of subjects.

The results are reported in Table 2 for the proposed functional mixture model and functional principal component analysis for different fixed values for the number of included components K. We find that the proposed method utilizes about 8 principal components on average ($K_{\text{avg}} = 8.27$) and with this number achieves better recovery, compared to the results obtained by the traditional functional principal component analysis using more components. Therefore, in this application, the proposed mixture model provides both better and more parsimonious fits.

Figure 2 displays egg-laying counts for 6 randomly selected flies, overlaid with smooth estimates obtained by the proposed mixture method and by traditional functional principal component analysis using 8 components (similar to K_{avg}) and also $K = 3$, a choice that explains 95% of the variation of the data and therefore would be adopted by the popular fraction of variance explained selection criterion. This figure indicates that the functional mixture method appears to adapt better to the varying shapes of the trajectories. The estimated probability densities of the first three mixture components and their mixture proportions are depicted in Figure 3.

Table 1: Average integrated squared error (AISE) and average number K_{avg} of principal components across all subjects. The first column denotes the type of data generation, either according to the mixture setting where the number of components varies from individual to individual, or according to the common setting, where the number of components is common for all subjects. The second column denotes the distribution of the number of principal components in the mixture setting and the number of common components in the common setting. The third column indicates the variance of the measurement error. The fifth and seventh columns show the AISE and the average number K_{avg} of chosen components for the proposed mixture model for the Gaussian process and non-Gaussian process, respectively, while these values are displayed in the sixth and eighth columns for functional principal component analysis (FPCA), along with the common choice K for the number of components. The Monte Carlo standard error based on 100 simulation runs is given in parentheses, multiplied by 100.

Simulation Setting				Gaussian		Non-Gaussian	
				MIPS	FPCA	MIPS	FPCA
mixture	multinomial	$\sigma^2 = 0.1$	AISE	$7.01_{(0.40)}$	$7.67_{(0.28)}$	$6.98_{(0.46)}$	$7.70_{(0.44)}$
			K_{avg}	$9.23_{(1.21)}$	$16.7_{(1.76)}$	$8.97_{(1.12)}$	$16.5_{(1.93)}$
		$\sigma^2 = 0.25$	AISE	$15.2_{(1.02)}$	$17.5_{(0.81)}$	$15.6_{(1.04)}$	$17.9_{(1.19)}$
			K_{avg}	$8.66_{(1.32)}$	$16.7_{(1.07)}$	$8.58_{(1.08)}$	$16.8_{(1.05)}$
	Poisson	$\sigma^2 = 0.1$	AISE	$5.63_{(0.21)}$	$6.32_{(0.23)}$	$5.82_{(0.65)}$	$6.61_{(0.89)}$
			K_{avg}	$6.78_{(0.14)}$	$13.7_{(1.13)}$	$6.68_{(0.27)}$	$13.4_{(1.43)}$
		$\sigma^2 = 0.25$	AISE	$12.1_{(0.37)}$	$13.9_{(0.32)}$	$12.2_{(0.66)}$	$14.0_{(1.05)}$
			K_{avg}	$6.63_{(0.16)}$	$14.5_{(1.17)}$	$6.28_{(0.23)}$	$13.4_{(1.94)}$
common	finite ($K = 12$)	$\sigma^2 = 0.1$	AISE	$6.55_{(0.07)}$	$6.46_{(0.07)}$	$6.56_{(0.07)}$	$6.46_{(0.07)}$
			K_{avg}	$13.6_{(0.43)}$	$12.1_{(0.47)}$	$13.5_{(0.50)}$	$12.2_{(0.56)}$
		$\sigma^2 = 0.25$	AISE	$15.7_{(0.19)}$	$15.5_{(0.15)}$	$15.8_{(0.23)}$	$15.5_{(0.21)}$
			K_{avg}	$12.9_{(1.00)}$	$12.6_{(1.03)}$	$12.7_{(0.81)}$	$12.6_{(0.99)}$
	infinite ($K = 25$)	$\sigma^2 = 0.1$	AISE	$13.2_{(0.01)}$	$12.9_{(0.01)}$	$13.3_{(0.12)}$	$12.9_{(0.14)}$
			K_{avg}	$24.3_{(0.05)}$	$25.0_{(0.00)}$	$24.1_{(0.09)}$	$25.0_{(0.00)}$
		$\sigma^2 = 0.25$	AISE	$32.0_{(0.53)}$	$31.4_{(0.51)}$	$31.9_{(0.38)}$	$31.5_{(0.70)}$
			K_{avg}	$23.8_{(0.19)}$	$25.0_{(0.00)}$	$23.6_{(0.17)}$	$25.0_{(0.00)}$

Table 2: Egg-laying data: 10-fold cross-validated relative squared errors (CVRSE), as obtained for the proposed functional mixture model and for traditional functional principal component analysis, where the latter uses a common number K of components across all subjects, for $K = 0, 2, \ldots, 18$ and K_{avg} is the mean of the number of principal components used by the proposed method. The results are based on 100 random partitions with standard error in parentheses.

FPCA	K	0	2	4	6	8
	CVRSE	$3.3914_{(.0070)}$	$.2038_{(.0003)}$	$.1549_{(.0002)}$	$.1388_{(.0002)}$	$.1347_{(.0001)}$
	K	10	12	14	16	18
	CVRSE	$.1340_{(.0001)}$	$.1337_{(.0001)}$	$.1336_{(.0001)}$	$.1335_{(.0001)}$	$.1334_{(.0001)}$
MIPS	CVRSE=$.1319_{(.0002)}$			$K_{\text{avg}} = 8.2684_{(.0192)}$		

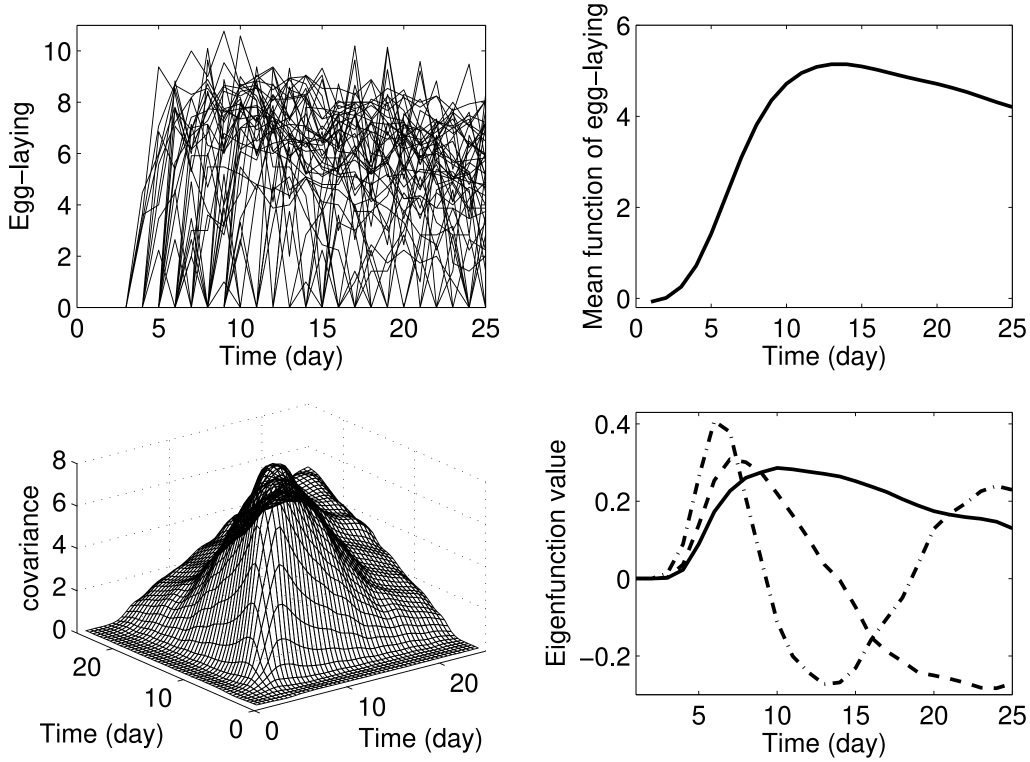

Figure 1: Top-left: Daily egg-laying counts, after square-root transformation, for 50 randomly selected flies for the first 25 days of their lifespan. Top-right: Smooth estimate of the mean function. Bottom-left: Smooth nonnegative definite estimate of the covariance surface. Bottom-right: Smooth estimates of the first (solid), second (dashed) and third (dash-dotted) eigenfunctions, explaining 95% of the total variation.

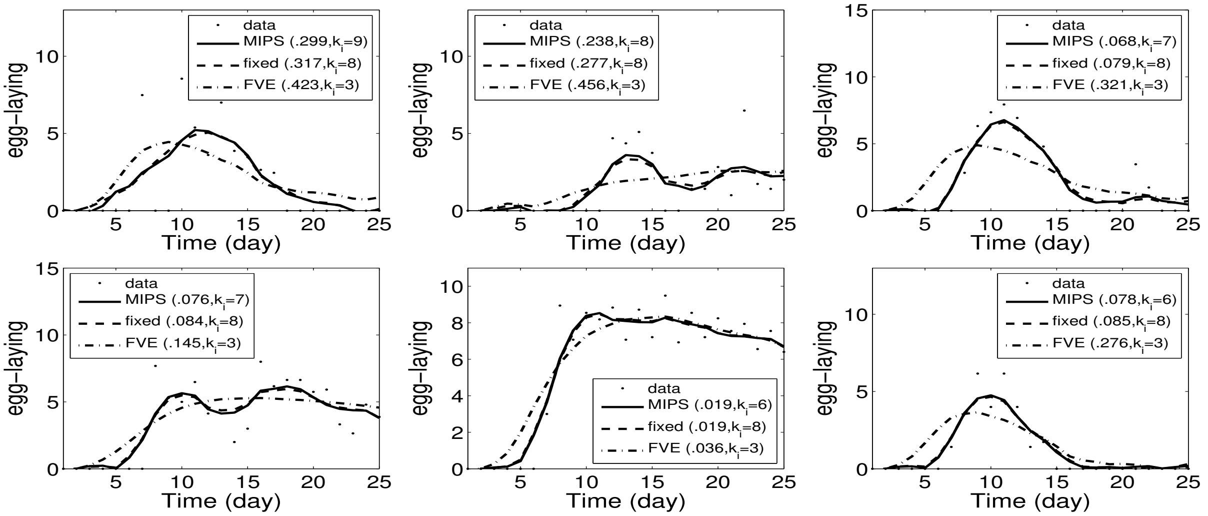

Figure 2: Daily egg-laying counts for six randomly selected flies for the first 25 days of their lifespan, overlaid with the smooth estimates obtained from the proposed functional mixture model, solid line, from functional principal component analysis with a fixed number of 8 components, dashed line, and for a fixed number of three components which explain 95% of the total variation, dash-dotted line. Shown in parentheses are the individual relative squared errors and the included number of components for that individual.

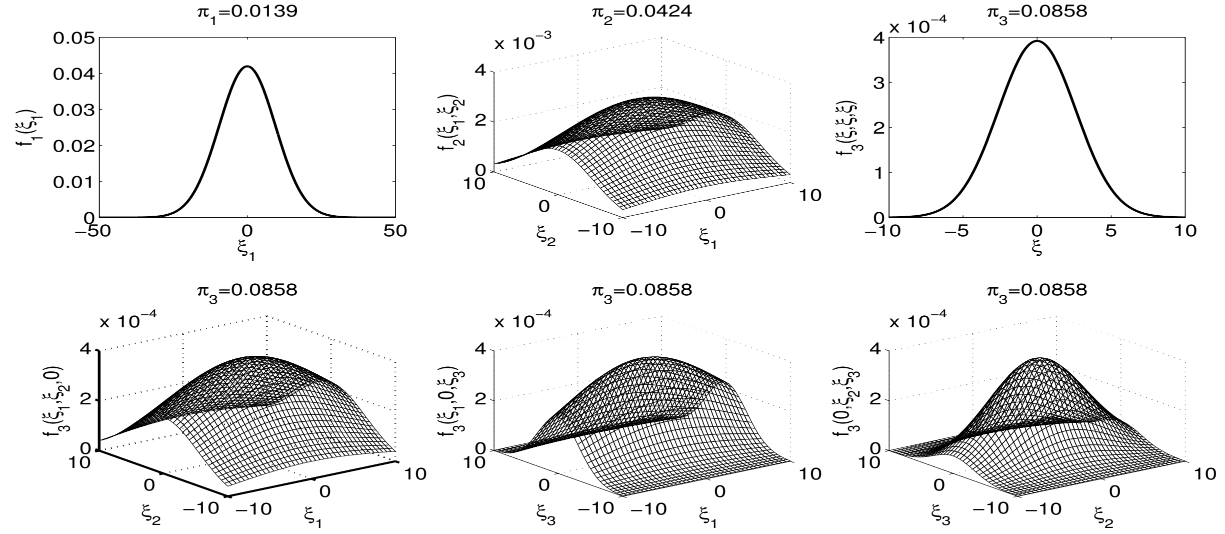

Figure 3: Estimated probability density function for the first three mixture components and their corresponding mixture proportions π_k. Top: left, the probability density function of the first component; middle, the probability density function of the second component; right, the diagonal of the probability density function of the third component. Bottom: 2-dimension slices of the probability densities of the third component at $\xi_1 = 0$, $\xi_2 = 0$ and $\xi_3 = 0$, respectively, in the left, middle and right panels.

6 Discussion and Concluding Remarks

Density functions are important for many statistical applications that require the construction of a likelihood, for example one could use maximum likelihood to find the best fit for a parametrized class of densities. Similarly, Bayes classifiers for functional data can be based on a density ratio. Densities can also be used for the estimation of modes and level contours and also to estimate the shape and location of ridges in functional data sets, extending the analogous problem for multivariate data. Specifically, the importance of constructing modes for functional data distributions has been recognized before (Gasser *et al.*, 1998). One can consider related mode finding algorithms that can be used for clustering functional data (Liu and Müller, 2003). As clustering of functional data is attracting increasing attention (Chiou and Li, 2007; Slaets *et al.*, 2012; Jacques and Preda, 2014) density based clustering for functional data likely will be of increasing interest for data analysis.

In addition to this relevance of densities in function space for functional data applications, the foundational issue of the existence and construction of densities in function space naturally puts the problem of obtaining a density for functional data into focus. The fact that such a density does not exist in the often considered function space L^2 demonstrates the scope of the problem. This has led to the construction of a surrogate density, which can be based on a truncated expansion of the functional data into functional principal components (Delaigle and Hall, 2010; Bongiorno and Goia, 2016). This construction is a workaround that provides a practical solution but leaves open the problem of finding a theoretical solution, for which one has to move away from the whole space L^2.

Motivated by practical consideration from applications of functional data analysis, we propose here a construction that provides a theoretical solution to the density problem by essentially considering random functions in L^2 whose distribution belongs to an infinite mixture of distributions on $k-$dimensional subspaces. Each of the component distributions has a finite dimension $k < \infty$ and corresponds to functions that can be fully described by an expansion into k components only. The space is still infinite-dimensional overall, as the dimensions k are unlimited. This mixture distribution approach has the advantage that an overall density can be well defined theoretically under regularity assumptions. Moreover, the components of the expansion can be estimated by applying a usual eigenexpansion that gives the correct eigenfunctions even if the mixture structure is ignored. To obtain the correct eigenvalues, the mixture probabilities play a role, and they can be consistently estimated under additional assumptions. We develop the construction of mixture inner product spaces for which appropriate mixture densities can be found under certain conditions in a framework of general infinite dimensional Hilbert spaces that transcends functional data analysis and therefore may be of more general interest. In data applications, the proposed mixture model tends to use fewer components than standard functional principal component analysis, while achieving the same or sometimes better approximations to the observed trajectories, which demonstrates that mixture inner product spaces are also of practical interest.

Acknowledgements

Z. Lin and F. Yao's research was partially supported by an individual discovery grant and DAS from NSERC, Canada. Z. Lin's research was also partially supported by Alexander Graham Bell Canada Graduate Scholarships from NSERC, Canada. H. G. Müller's research was partially supported by NSF grant DMS-1407852, U.S.A..

Appendix: Technical Proofs

Proof of Proposition 1. Let x be an arbitrary element of H and $a_k = \langle x, \phi_k \rangle$. Since $\phi_1, \phi_2, \ldots$ form a complete orthonormal basis of H, we have $\|x\|^2 = \sum_{k=1}^{\infty} a_k^2 < \infty$. Now define $x_k = \sum_{j=1}^{k} a_j \phi_k$. Then $x_k \in S$ for each $k = 1, 2, \ldots$. Also, $\|x - x_k\|^2 = \sum_{j=k+1}^{\infty} a_j^2 \to 0$ as $k \to \infty$. This implies that for any $h > 0$, the open ball $B(x; h)$ with center at x and radius h contains some $x_k \in S$ for some k. This shows that S is dense in H.

To show part (2), note that $H_k = \bigcup_{j=1}^{k} S_j$ and hence $S = \bigcup_{k=1}^{\infty} S_k = \bigcup_{k=1}^{\infty} \bigcup_{j=1}^{k} S_j = \bigcup_{k=1}^{\infty} H_k$. Since each H_k is a closed subset of H and hence $H_k \in \mathscr{B}(H)$, we conclude that $S = \bigcup_{k=1}^{\infty} H_k$ is in $\mathscr{B}(H)$. To see $\mathscr{B}(S) \subset \mathscr{B}(H)$, we first note that, since the metric d_S on S, defined by $d_S(x,y) = \|x - y\|_H$ for all $x, y \in S \subset H$, is the restriction of the metric d_H on H, the subspace topology of S coincides with the topology induced by the metric d_S. This implies that for any open set A of S there exists an open subset B of H such that $A = B \cap S$. As both B and S are in $\mathscr{B}(H)$, we have $A \in \mathscr{B}(H)$. In other words, the collection τ_S of all open sets of S is a subset of $\mathscr{B}(H)$. This implies $\mathscr{B}(S) \subset \mathscr{B}(H)$, recalling that $\mathscr{B}(S)$ is the smallest σ-algebra containing τ_S.

For part (3), we first note that $\mathscr{B}(S) = \{B \cap S : B \in \mathscr{B}(H)\}$, by Lemma 3 in Chapter II of Shiryaev (1984). Now, if $B \in \mathscr{B}(H)$, then $B \cap S \in \mathscr{B}(S)$ and hence $X_S^{-1}(B) = X_S^{-1}(B \cap S) \in \mathscr{E}$. Therefore, X_S is also $\mathscr{E}$-$\mathscr{B}(H)$ measurable and hence an H-valued random element. □

Proof of Proposition 2. We prove the claim by explicitly constructing such an S-valued random element Y, as follows. Let $\varepsilon_1 = \{E(\|X - X_k\|_H^p)\}^{1/p}$ and $\delta = (\varepsilon - \varepsilon_1)/2 > 0$. Since $f_k(0) > 0$ and f_k is continuous at 0, if $\Omega_\delta = \{\omega \in \Omega : \xi_k(\omega) \in (-\delta/2, \delta/2)\}$, then $P(\Omega_\delta) > 0$. Define $Y(\omega) = X_k(\omega)$ if $\omega \notin \Omega_\delta$ and $Y(\omega) = X_{k-1}(\omega)$ otherwise. If we define $Z(\omega) = \xi_k(\omega)\phi_k 1_{\Omega_\delta}$, then $Y = X_k - Z$. Since $\{E(\|Z\|_H^p)\}^{1/p}$ defines a norm on all H-valued random elements Z such that $\{E(\|Z\|_H^p)\}^{1/p} < \infty$ (Vakhania *et al.*, 1987), this implies that $\{E(\|X - Y\|_H^p)\}^{1/p} = \{E(\|X - X_k + Z\|_H^p)\}^{1/p} \leq \{E(\|X - X_k\|_H^p)\}^{1/p} + \{E(\|Z\|_H^p)\}^{1/p} < \varepsilon_1 + \delta P(\Omega_\delta) < \varepsilon$. On the other hand, the continuity of f_k at 0 implies that $P(\xi_k = 0) = 0$, and hence we have $\mathscr{K}(Y) = \mathscr{K}(X_k) - P(\Omega_\delta) < \mathscr{K}(X_k)$. □

Proof of Theorem 2. Note that each Lebesgue measure τ_k is σ-finite. This means that for each k there is a countable partition $S_{k1}, S_{k2,} \ldots$ of S_k such that $S_{kj} \in \mathscr{B}(S)$ and $\tau_k(S_{kj}) < \infty$ for all $j = 1, 2, \ldots$. Since $S = \bigcup_k \bigcup_j S_{kj}$, we know that $\{S_{kj} : j = 1, \ldots, k = 1, \ldots\}$ forms a countable partition of S, where each S_{kj} has finite measure $\tau(S_{kj}) = \tau_k(S_{kj}) < \infty$. This shows that τ is σ-finite.

To show that P_X is absolutely continuous to τ, suppose $A \in \mathscr{B}(S)$ and $\tau(A) = 0$, and define $A_k = A \cap S_k$. Then $\tau_k(A_k) = 0$ for all k. Note that $P_X(A) = \sum_{k=1}^{\infty} P_X(A_k)$. Define $\eta_k(x) = (\langle x, \phi_1 \rangle, \langle x, \phi_2 \rangle, \ldots, \langle x, \phi_k \rangle) \in \mathbb{R}^k$ for each $x \in H_k$. Note that each η_k is a canonical isomorphic mapping between H_k and $\mathbb{R}^k$. Thus, the Lebesgue measure of $\eta_k(A_k)$ is equal to $\tau_k(A_k)$ and is zero. Now, $P_X(A_k) = P\{(\xi_1, \xi_2, \ldots, \xi_k) \in \eta_k(A_k), X = k\} = P\{(\xi_1, \xi_2, \ldots, \xi_k) \in \eta_k(A_k) \mid X = k\} P(X = k) =$
$\pi_k \int_{\eta_k(A_k)} f_k(t_1, t_2, \ldots, t_k) dt_1 dt_2 \cdots dt_k = 0$, where the last equality is due the fact that the Lebesgue measure of $\eta_k(A_k)$ is zero and the fact that f_k is a density function by assumption. Therefore, $P_X(A) = 0$, and we conclude that P_X is absolutely continuous w.r. to τ.

By the Radon-Nykodym theorem, there is a density f of P_X on S with respect to τ. Now we show that f defined in (1) is such a density. Let $A \in \mathscr{B}(S)$. As above we define $A_k = A \cap S_k$. Then $A_1, A_2, \ldots$ form a partition of A, and hence

$$\int_A f \mathrm{d}\tau = \sum_k \int_{A_k} f \mathrm{d}\tau = \sum_k \pi_k \int_{A_k} f_k \mathrm{d}\tau_k. \tag{14}$$

Now, for each k,

$$\begin{aligned} P_X(A_k) &= \Pr\{(\xi_1, \xi_2, \ldots, \xi_k) \in \eta(A_k), K = k\} = \pi_k \Pr\{(\xi_1, \xi_2, \ldots, \xi_k) \in \eta(A_k) \mid K = k\} \\ &= \pi_k \int_{\eta(A_k)} f_k(t_1, t_2, \ldots, t_k) dt_1 dt_2 \cdots dt_k = \pi_k \int_{A_k} f_k d\tau_k. \end{aligned} \tag{15}$$

Given (14) and (15), we conclude that $\int_A f \mathrm{d}\tau = \sum_k P_X(A_k) = P_X(A)$, and hence f is a probability density function of P_X w.r. to τ. □

To simplify notations, we simply use r from now on, while one should be aware that r grows to infinity as sample size $n \to \infty$. The proof of Theorem 3 requires several lemmas.

Let $\hat{G}(s,t) = n^{-1} \sum_{i=1}^{n} X_i(s) X_i(t)$ denote the empirical version of $G(s,t)$ and $\hat{\phi}_k$ be the kth eigenfunction of $\hat{G}$. When it is clear from the context, we use G and $\hat{G}$ to denote the corresponding covariance operator. Define $\hat{\Delta} = \{\int_{D \times D} (\hat{G}(s,t) - G(s,t))^2 ds dt\}^{1/2}$ and for a constant $C_4 > 0$,

$$J' = \{j - 1 : \lambda_j - \lambda_{j+1} \geq 2\hat{\Delta}\}, \quad \text{and} \quad J = \{j \in J' : j \leq C_4 n^{1/(2b+2)}\}. \tag{16}$$

From $\hat{\Delta} = O_p(n^{-1/2})$ (Hall and Hosseini-Nasab, 2006) and assumption (A3), we have

$$P(C_5 n^{1/(2b+2)} \leq \sup J \leq C_4 n^{1/(2b+2)}) \to 1$$

for a positive constant $C_5 \leq C_4$. The following lemma quantifies the estimation quality of the eigenfunctions $\hat{\phi}_k$ and the principal component scores $\hat{\xi}_{ik}$. Let $\hat{\xi}_{i,(k)} = (\hat{\xi}_{i,1}, \hat{\xi}_{i,2}, \ldots, \hat{\xi}_{i,k})$ where $\hat{\xi}_{ij} = \langle X_i, \hat{\phi}_j \rangle$.

Lemma 1. *If assumptions (A0), (A2) and (A3) hold,*

$$\sup_{n\geq 1}\sup_{k\in J} n^2 k^{-4(b+1)} E\|\hat{\phi}_k - \phi_k\|^4 < \infty; \tag{17}$$

$$\sup_{n\geq 1}\sup_{k\in J} n k^{-2b-3} E\|\xi_{i,(k)} - \hat{\xi}_{i,(k)}\|^2 < \infty. \tag{18}$$

Proof. The bound in (17) directly follows from Lemma 3.4 of Hall and Hosseini-Nasab (2009), $E(\hat{\Delta}^4) = O(n^{-2})$ (Lemma 3.3 of Hall and Hosseini-Nasab, 2009) and (A3). To show (18),

$$\begin{aligned} E\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^2 &= \sum_{j=1}^{k} E(|\hat{\xi}_{i,j} - \xi_{i,j}|^2) = \sum_{j=1}^{k} E(\langle X_i, \hat{\phi}_j - \phi_j\rangle^2) \\ &\leq \sum_{j=1}^{k} E(\|X_i\|^2 \|\hat{\phi}_j - \phi_j\|^2) \leq \sum_{j=1}^{k} \{E(\|X_i\|^4) E(\|\hat{\phi}_j - \phi_j\|^4)\}^{1/2} \\ &= \{E(\|X\|^4)\}^{1/2} \sum_{j=1}^{k} \{E(\|\hat{\phi}_j - \phi_j\|^4)\}^{1/2}. \end{aligned}$$

Then (18) follows with the fact $E(\|X\|^4) < \infty$ and (17). □

We next examine the discrepancy between true and estimated likelihood functions. Recall that $Q_r = \min(K, r)$, $Z = \sum_{j=1}^{Q_r} \langle X, \phi_j\rangle \phi_j$, the log-likelihood of Z with $\pi_r^* = 1 - \sum_{k=1}^{r-1} \pi_k$,

$$L_{r,1}(Z \mid \theta_{[r]}) = \log\left\{\pi_r^* f_r(Z \mid \theta_{(r)}) 1_{Z\in S_r} + \sum_{k=1}^{r-1} \pi_k f_k(Z \mid \theta_{(k)}) 1_{Z\in S_k}\right\},$$

and $L_r(\theta_{[r]}) = E\{L_{r,1}(z \mid \theta_{[r]})\}$. Define the log-likelihood function of θ given $Z_1, \ldots, Z_n$ by $L_{r,n}(\theta_{[r]}) = \frac{1}{n}\sum_{i=1}^{n} L_{r,1}(Z_i \mid \theta_{[r]})$. The following lemma quantifies the discrepancy between $L_{r,n}(\theta_{[r]})$ and $L_r(\theta_{[r]})$.

Lemma 2. *If the assumptions in Theorem 3 hold, then for each* $\theta_{[r]}$,

$$r^a |L_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]})| \xrightarrow{p} 0.$$

Proof. We first express $L_r(\theta_{[r]}) = E\{L_{r,1}(z \mid \theta_{[r]})\}$ as follows,

$$\begin{aligned}
L_r(\theta_{[r]}) &= E\log\left\{\pi_r^* f_r(Z \mid \theta_{(r)}) 1_{Z\in S_r} + \sum_{k=1}^{r-1} \pi_k f_k(Z_i \mid \theta_{(k)}) 1_{Z\in S_k}\right\} \\
&= EE\left[\log\{\pi_r^* f_r(Z \mid \theta_{(r)}) 1_{Z\in S_r} + \sum_{k=1}^{r-1} \pi_k f_k(Z \mid \theta_{(k)}) 1_{Z\in S_k}\} \mid Q_r\right] \\
&= \pi_r^* E\left[\log\{\pi_r^* f_r(Z \mid \theta_{(r)})\} \mid K \geq r\right] + \sum_{k=1}^{r-1} \pi_k E\left[\log\{\pi_k f_k(Z \mid \theta_{(k)})\} \mid K = k\right] \qquad (19) \\
&= \pi_r^* \log \pi_r^* + \sum_{k=1}^{r-1} \pi_k \log \pi_k + \pi_r^* E\{g_r(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_r)\} + \sum_{k=1}^{r-1} \pi_k E\{g_k(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)\},
\end{aligned}$$

where (19) is obtained by noting that $1_{Z\in S_k} = 0$ if $Q_r \neq k$, and $1_{Z\in S_k} = 1$ if $Q_r = k$. Let $W_{n,i} = L_{r,1}(Z_i \mid \theta_{[r]}) - L_r(\theta_{[r]})$. Then $E(W_{n,i}) = 0$ and $L_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]}) = n^{-1}\sum_{i=1}^n W_{n,i}$. We shall show that $r^a W_{n,i}$ satisfies the Cesàro type uniform integrability defined in Sung (1999), and hence admits a weak law of large numbers. First, we show that $\sup_{n\geq 1} n^{-1}\sum_{i=1}^n r^a E|W_{n,i}| = O(1)$. For sufficiently large n, given (A6), with definition $\tilde{\xi}_{(k)} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_k)$, we have

$$\begin{aligned}
E|W_{n,i}| &\leq E|g_r(Z) 1_{Z\in S_r} - \pi_r^* E(g_r(\tilde{\xi}_{(r)}))| + \sum_{k=1}^{r-1} E|g_k(Z) 1_{Z\in S_k} - \pi_k E(g_k(\tilde{\xi}_{(k)}))| \\
&= E|g_r(\tilde{\xi}_{(r)}) 1_{K\geq r} - \pi_r^* E(g_r(\tilde{\xi}_{(r)}))| + \sum_{k=1}^{r-1} E|g_k(\tilde{\xi}_{(k)}) 1_{K=k} - \pi_k E(g_k(\tilde{\xi}_{(k)}))| \\
&\leq 2\pi_r^* E|g_r(\tilde{\xi}_{(r)})| + 2\sum_{k=1}^{r-1} \pi_k E|g_k(\tilde{\xi}_{(k)})| \leq r^{-a+c-\beta+1} + c_2 \sum_{k=1}^{r-1} k^{-a+c-\beta} \\
&\leq c_3 r^{-a}, \qquad (20)
\end{aligned}$$

where c_3 is a constant that does not depend on n and i, and the last inequality is due to the condition $\beta > c+1$ in (A6). Thus $\sup_{n\geq 1} n^{-1}\sum_{i=1}^n E\{r^a|W_{n,i}|\} = O(1)$. Since for any n, the random variables $W_{n,1}, W_{n,2}, \ldots, W_{n,n}$ are i.i.d.,

$$\lim_{u\to\infty} \sup_{n\geq 1} n^{-1} \sum_{i=1}^n E\{r^a |W_{n,i}| 1_{|W_{n,i}|>u}\} = \lim_{u\to\infty} \sup_{n\geq 1} E\{r^a |W_{n,1}| 1_{|W_{n,1}|>u}\} = 0, \qquad (21)$$

as $E\{r^a|W_{n,1}|\} < \infty$ by (20). Then $r^a W_{n,i}$ satisfies the Cesàro type uniform integrability, and the conclusion of the lemma follows from a weak law of large numbers for triangular arrays (Sung, 1999). □

We are now ready to quantify the discrepancy from the estimated likelihood function $\hat{L}_{r,n}$ by plugging in the estimated quantities $\hat{\phi}_k$ and $\hat{\tilde{\xi}}_{ik}$.

Lemma 3. *If the assumptions in Theorem 3 hold, then for each* $\theta_{[r]}$,

$$r^a|\hat{L}_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]})| \xrightarrow{p} 0.$$

Proof. Recall $\hat{\xi}_{i,(k)} = (\hat{\xi}_{i,1}, \hat{\xi}_{i,2}, \ldots, \hat{\xi}_{i,k})$, and define

$$\begin{aligned} Y_{n,i} &= \pi_r^*\{C_1 H_r(\xi_{i,(r)})\|\hat{\xi}_{i,(r)} - \xi_{i,(r)}\|^{\nu_1} 1_{Z_i \in S_r} + C_2 r^{\alpha_2}\|\hat{\xi}_{i,(r)} - \xi_{i,(r)}\|^{\nu_2} 1_{Z_i \in S_r}\} \\ &\quad + \sum_{k=1}^{r-1} \pi_k\{C_1 H_k(\xi_{i,(k)})\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^{\nu_1} 1_{Z_i \in S_k} + C_2 k^{\alpha_2}\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^{\nu_2} 1_{Z_i \in S_k}\}. \end{aligned}$$

By (A4), we have

$$|\hat{L}_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]})| \leq |L_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]})| + n^{-1}\sum_{i=1}^{n} Y_{n,i}.$$

From the condition $r = n^{\gamma-\varepsilon}$ in Theorem 3 and the definition of J in (16), we have $P\{r \in J\} \to 1$ as $n \to \infty$. Thus we may assume $r \in J$ in the sequel. With Lemma 1, if $\nu' \leq 2$, then $E\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^{\nu'} \leq \{E\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^2\}^{\nu'/2} = O(k^{(2b+3)\nu'/2} n^{-\nu'/2})$ uniformly for $k \leq r$ and n. Since $2\nu_1 \leq 2$, $\nu_2 \leq 2$, $\alpha = \max(\alpha_1, \alpha_2)$ and $\nu = \min(2\nu_1, \nu_2)$, for some $c_0 > 0$,

$$\begin{aligned} E\{|H_k(\xi_{i,(k)})|\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^{\nu_1}\} &\leq [E\{H_k(\xi_{i,(k)})\}^2]^{1/2}(E\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^{2\nu_1})^{1/2} \\ &= O\big(k^{(2b+3)\nu/2+\alpha} n^{-\nu/2}\big), \\ E\{k^{\alpha_2}\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^{\nu_2}\} &= O\big(k^{(2b+3)\nu/2+\alpha} n^{-\nu/2}\big). \end{aligned}$$

Recall $\gamma_1 = (2b+3)\nu/2 + \alpha - 2\beta$, $\gamma_2 = a + (\gamma_1 + 2)1_{\gamma_1 > -2}$, $\gamma = \min\{\nu/(2\gamma_2), 1/(2b+2)\}$ in Theorem 3, implying $\gamma\gamma_2 \leq \nu/2$, and hence

$$\begin{aligned} E|Y_{n,i}| &\leq (\pi_r^*)^2 E\{C_1|H_r(\xi_{i,(r)})|\|\hat{\xi}_{i,(r)} - \xi_{i,(r)}\|^{\nu_1} + C_2 r^{\alpha_2}\|\hat{\xi}_{i,(r)} - \xi_{i,(r)}\|^{\nu_2}\} \\ &\quad + \sum_{k=1}^{r-1} \pi_k^2 E\{C_1|H_k(\xi_{i,(k)})|\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^{\nu_1} + C_2 k^{\alpha_2}\|\hat{\xi}_{i,(k)} - \xi_{i,(k)}\|^{\nu_2}\} \\ &\leq c_1 r^{-2\beta+2} r^{(2b+3)\nu/2+\alpha} n^{-\nu/2} + c_2 \sum_{k=1}^{r-1} k^{-2\beta+(2b+3)\nu/2+\alpha} n^{-\nu/2} \\ &= c_1 r^{\gamma_1+2} n^{-\nu/2} + c_2 \sum_{k=1}^{r-1} k^{\gamma_1} n^{-\nu/2} \leq c_3 n^{-\nu/2} r^{\gamma_2 - a} \leq c_4 r^{-a} n^{-\varepsilon\gamma_2/2}, \end{aligned} \tag{22}$$

where the last inequality is due to $r = O(n^{\gamma-\varepsilon})$, and c_1, c_2, c_3, c_4 are positive constants that do not depend on n. Setting $\delta = 3/(3 + \varepsilon\nu/\gamma_2) < 1$, by the Lyapunov inequality, $r^{a\delta}E|Y_{n,i}|^\delta \leq r^{a\delta}(E|Y_{n,i}|)^\delta \leq c_4 r^{a\delta} r^{-a\delta} n^{-\delta\varepsilon\nu/(2\gamma_2)} = c_4 n^{-\delta\varepsilon\nu/(2\gamma_2)}$ uniformly for n and $r = O(n^{\gamma-\varepsilon})$. Although the $Y_{n,i}$ are not independent of $Y_{n,j}$, they have the same distribution due to symmetry. Therefore, noting that $1 - \delta\{1 +$

$\varepsilon\nu/(2\gamma_2)\} < 0$, we have

$$\sup_{n\geq 1} n^{-\delta}\sum_{i=1}^{n} E\{r^{a\delta}|Y_{n,i}|^{\delta}\} \leq \sup_{n\geq 1} c_4 n^{-\delta} n n^{-\delta\varepsilon\nu/(2\gamma_2)} = \sup_{n\geq 1} c_4 n^{1-\delta\{1+\varepsilon\nu/(2\gamma_2)\}} = O(1). \tag{23}$$

The result $r^{a\delta}E|Y_{n,i}|^{\delta} = O(n^{-\delta\varepsilon\nu/(2\gamma_2)})$ also implies that

$$\lim_{M\to\infty}\sup_{n\geq 1} n^{-\delta}\sum_{i=1}^{n} E\{r^{a\delta}|Y_{n,i}|^{\delta} 1_{|Y_{n,i}|^{\delta}>M}\} = 0. \tag{24}$$

Then the Cesàro type uniform integrability is satisfied by $r^a Y_{n,i}$ with exponent $\delta < 1$, based on (23) and (24), and the weak law of large numbers (Sung, 1999) implies $n^{-1}\sum_{i=1}^{n} r^a Y_{n,i} = o_p(1)$. This result, in conjunction with the fact $r^a|L_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]})| = o_p(1)$ and $r^a|\hat{L}_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]})| \leq r^a|L_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]})| + n^{-1}\sum_{i=1}^{n} r^a Y_{n,i}$, as well as $\Pr\{r \in J\} \to 1$, yields the result. □

Proof of Theorem 3. By assumption (A5), $\theta_{[r],0}$ is the maximizer of $L_r(\theta_{[r]})$. Let $h_5 = \min\{h_1, h_2, h_3\}$ and $U_r^a = \{\theta_{[r]} : L_r(\theta_{[r],0}) - L_r(\theta_{[r]}) < h_5 r^{-a}\}$, whence $U_r^a \subset U_r \subset \mathcal{B}_r$, where $a = \max(a_1, a_2)$, U_r and $\mathcal{B}_r$ are defined in (A5). Moreover, for all $\theta_{[r]} \in U_r^a$, there exists $h_4 > 0$ not depending on r and $\theta_{[r]}$, such that

$$L_r(\theta_{[r],0}) - L_r(\theta_{[r]}) \geq h_4 r^{-a}\|\theta_{[r]} - \theta_{[r],0}\|^2, \tag{25}$$

From (A1), $\Theta = \prod_{j=1}^{\infty} I_{[\infty],j}$ is compact due to Tychonoff's theorem, which implies that the convergence of $r^a|\hat{L}_{r,n}(\theta_{[r]}) - L_r(\theta_{[r]})|$ in Lemma 3 is uniform over Θ. Thus for any $0 < \varepsilon^2 < h_5$, there exists $N_\varepsilon > 0$ such that if $n > N_\varepsilon$, then

$$\Pr\Big(\{r^a|\hat{L}_{r,n}(\theta_{[r],0}) - L_r(\theta_{[r],0})| < \varepsilon^2/2\}\bigcap\{r^a|\hat{L}_{r,n}(\hat{\theta}_{[r]}) - L_r(\hat{\theta}_{[r]})| < \varepsilon^2/2\}\Big) > 1 - \varepsilon/2, \tag{26}$$

where $\hat{\theta}_{[r]}$ is a global maximizer of $\hat{L}_{r,n}$. Next we show that

$$\Pr\{r^a|\hat{L}_{r,n}(\hat{\theta}_{[r]}) - L_r(\theta_{[r],0})| < \varepsilon^2/2\} > 1 - \varepsilon/2. \tag{27}$$

If $\hat{L}_{r,n}(\hat{\theta}_{[r]}) \geq L_r(\theta_{[r],0})$, then $0 \leq \hat{L}_{r,n}(\hat{\theta}_{[r]}) - L_r(\theta_{[r],0}) \leq \hat{L}_{r,n}(\hat{\theta}_{[r]}) - L_r(\hat{\theta}_{[r]})$ since $L_r(\hat{\theta}_{[r]}) \leq L_r(\theta_{[r],0})$, due to the fact that $\theta_{[r],0}$ is the global maximizer of $L_r(\cdot)$. Similarly, if $\hat{L}_{r,n}(\hat{\theta}_{[r]}) \leq L_r(\theta_{[r],0})$, then $0 \leq L_r(\theta_{[r],0}) - \hat{L}_{r,n}(\hat{\theta}_{[r]}) \leq L_r(\theta_{[r],0}) - \hat{L}_{r,n}(\theta_{[r],0})$ since $\hat{L}_{r,n}(\theta_{[r],0}) \leq \hat{L}_{r,n}(\hat{\theta}_{[r]})$ due to the fact that $\hat{\theta}_{[r]}$ is a global maximizer of $\hat{L}_{r,n}(\cdot)$. Combining these two cases yields $|\hat{L}_{r,n}(\hat{\theta}_{[r]}) - L_r(\theta_{[r],0})| \leq \max\{|\hat{L}_{r,n}(\hat{\theta}_{[r]}) - L_r(\hat{\theta}_{[r]})|, |L_r(\theta_{[r],0}) - \hat{L}_{r,n}(\theta_{[r],0})|\}$. This result, in conjunction with (26), yields (27).

Then applying the triangle inequality in conjunction with (26) and (27) leads to $\Pr\{r^a|L_r(\hat{\theta}_{[r]}) - L_r(\theta_{[r],0})| < \varepsilon^2\} > 1 - \varepsilon$. Since $\varepsilon^2 < h_5$, we have $\hat{\theta}_{[r]} \in U_r^a$ with probability $(1-\varepsilon)$, and then apply (25) to conclude that $\Pr\{\|\hat{\theta}_{[r]} - \theta_{[r],0}\| < 2\varepsilon/\sqrt{h_4}\} > 1 - \varepsilon$, which yields the consistency of $\hat{\theta}_{[r]}$.

It remains to show the consistency of $f(x \mid \hat{\theta}_{[r]})$ for any $x \in \bigcup_{k=1}^{\infty} S_k$, which implies that there exists some $k_0 < \infty$ such that $x \in S_{k_0}$. Then $f(x) = \sum_{k=1}^{k_0} f_k(x \mid \theta_k) 1_{S_k}$, as the indicator functions 1_{S_j} are all zero if $j > k_0$. For sufficiently large n such that $k_0 \leq r_n$, $\theta_{[r_n]}$, and hence $\theta_1, \ldots, \theta_{k_0}$ are all consistently estimated. The continuity of each f_k with respect to θ_k in (A4) then implies that $\left| f(x \mid \hat{\theta}_{[r]}) - f(x \mid \theta_{[\infty],0}) \right| \xrightarrow{p} 0$.

□

References

Benaglia, T., Chauveau, D. and Hunter, D. R. (2009). An em-like algorithm for semi- and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics* **18** 505–526.

Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* **51** 285–311.

Boente, G. and Fraiman, R. (2000). Kernel-based functional principal components. *Statistics & Probability Letters* **48** 335–345.

Bongiorno, E. G. and Goia, A. (2016). Some insights about the small ball probability factorization for Hilbert random elements. *arXiv:1501.04308v2* preprint.

Carey, J. R., Liedo, P., Müller, H.-G., Wang, J.-L. and Chiou, J.-M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large, cohort of mediterranean fruit fly females. *Journal of Gerontology - Biological Sciences and Medical Sciences* **54** B245–251.

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.

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

Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 679–699.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis* **12** 136–154.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics* **38** 1171–1193.

Gasser, T., Hall, P. and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 681–691.

GIKHMAN, I. I. and SKOROKHOD, A. V. (1969). *Introduction to the Theory of Random Processes*. W.B. Saunders.

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. *The Annals of Statistics* **35** 70–91.

HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 109–126.

HALL, P. and HOSSEINI-NASAB, M. (2009). Theory for high-order bounds in functional principal components analysis. *Mathematical Proceedings of the Cambridge Philosophical Society* **146** 225–256.

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

HALL, P. and VIAL, C. (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society B* **68** 689–705.

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley.

JACQUES, J. and PREDA, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* **71** 92–106.

KNEIP, A. and UTIKAL, K. J. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* **96** 519–542.

LEVINE, M., HUNTER, D. R. and CHAUVEAU, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98** 403–416.

LI, W. V. and LINDE, W. (1999). Approximation, metric entropy and small ball estimates for Gaussian measures. *The Annals of Probability* **27** 1556–1578.

LI, Y. and GUAN, Y. (2014). Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance. *Journal of the American Statistical Association* **109** 1205–1215.

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

LI, Y., WANG, N. and CARROLL, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association* **108** 1284–1294.

LIU, X. and MÜLLER, H.-G. (2003). Modes and clustering for time-warped gene expression profile data. *Bioinformatics* **19** 1937–1944.

NIANG, S. (2002). Estimation de la densité dans un espace de dimension infinie: Application aux diffusions. *Comptes Rendus Mathematique* **334** 213–216.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics, 2nd edition. Springer, New York.

RAO, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics* **14** 1–17.

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B* **53** 233–243.

SHIRYAEV, A. N. (1984). *Probability*. Vol. 95 of Graduate Texts in Mathematics. Springer-Verlag, New York.

SILVERMAN, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* **24** 1–24.

SLAETS, L., CLAESKENS, G. and HUBERT, M. (2012). Phase and amplitude-based clustering for functional data. *Computational Statistics & Data Analysis* **56** 2360–2374.

SUNG, S. H. (1999). Weak law of large numbers for arrays of random variables. *Statistics & Probability Letters* **42** 293–298.

VAKHANIA, N. N., TARIELADZE, V. I. and CHOBANYAN, S. A. (1987). *Probability Distributions on Banach Spaces*. Reidel Publishing, Dordrecht.

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.

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