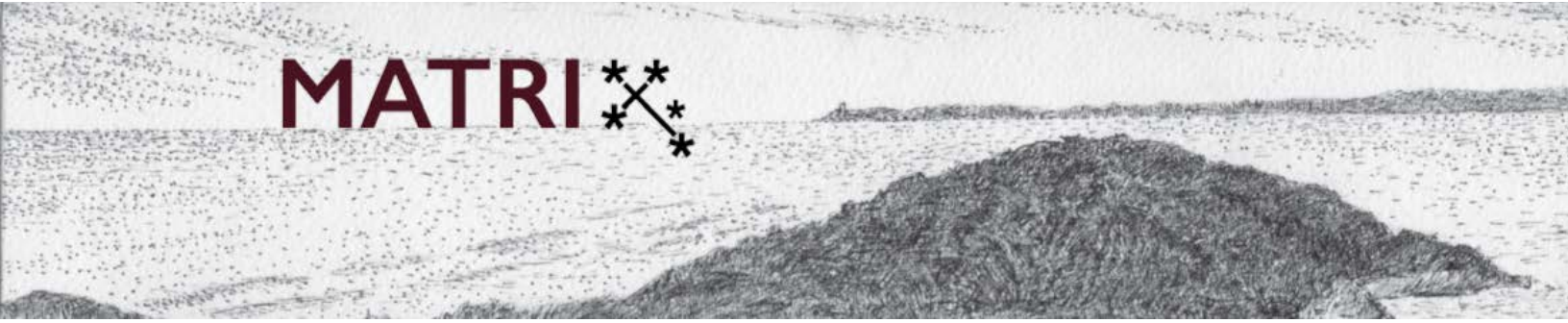
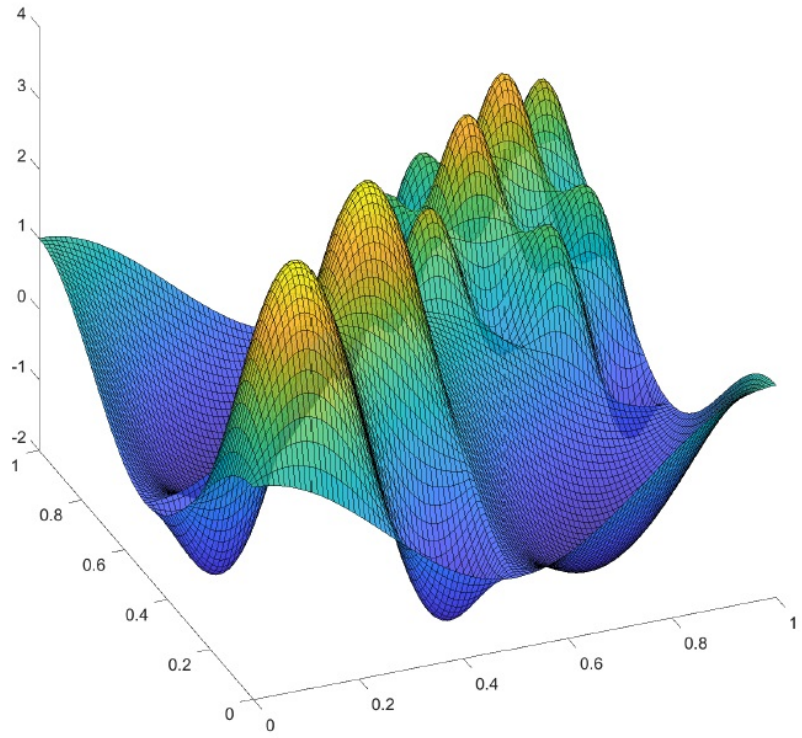


Functional Data Analysis and Beyond

Workshop



Talk Title and Abstracts

MATRIX, Creswick, 3 – 14 December 2018

FDA Workshop in MATRIX: Title and abstract

Speaker: Michelle Carey (University College Dublin)

Title: Dynamic modelling for Geospatial Processes
Abstract: Geo-spatial statistical analysis characterises processes evolving over time and over complex spatial regions in the presence of uncertain, incomplete and often noisy observations. The tradition in statistics, derived from time series analysis, has been to express these models descriptively in terms of first and second moments. However, most real-world processes are more complex than can be specified by the relatively simple classes of covariance functions and as such these processes may be more concisely described in terms of a dynamical model that considers the evolution explicitly using a partial differential equation (PDE) as advocated by Ramsay (2002), Sangalli et al (2013), Zhou and Pan (2014). Recently, hybrid models have been developed that utilise a stochastic partial differential equation (SPDE) to attain a more complex covariance function that can be subsequently incorporated into the descriptive modelling approaches (Lindgren et al (2011), Fuglstad et al (2015)). Here we review and compare the hybrid and PDE approaches noting recent progress, and identifying important challenges for the future
Joint work with James O. Ramsay.

Speaker: Sophie Dabo-Niang (Universite Lille 3)

Title: Bridging FDA, cell biomechanical phenotype and their biological expressions for Cancer diagnosis

Abstract: Most cancer patients die due to metastasis, and the early onset of this multistep process is usually missed by current staging tumour modalities [1]. Advanced technics exist to enrich disseminated tumour cells from patient blood and bone mallow as cancer progression marker. However, these cells present high heterogeneity [2], only some of them can exhibit stem cell phenotype and tumour development potential, other can have plasticity potential to reprogram into cancer stem cells [3]. So, detection and characterization is challenging due to lack of clear phenotypic markers [4]. Therefore, there is a critical need to find new ways to anticipate and predict metastasis development at an early stage of patient care.
Cancer progression involves many cellular morphological effects, which have been revealed by biophysical studies [5]. The relevance of the characterization of cancer cells by their electromechanical phenotype is attested by reports pointing out their physical alteration as reduced deformability of circulating lymphocytes in the case of chronic lymphocytic leukaemia, and increased size and stiffness of labelled circulating tumor cells (CTC). The physical properties allow identifying different malignant breast epithelial cells lines by their viscoelastic behavior [6] or their electrical impedance [7].

Even though, the analysis capability of cancer cell by their physical characteristics has been demonstrated [5], the current state of the art is far from being relevant for clinical practices. This project aims to develop and push the concept of cell physical phenotyping to categorize disseminating cells, evaluate their metastatic potential towards cancer diagnosis and prognosis.

We aim to combine MEMS (Micro-Electro-Mechanical Systems), biological assays, statistical learning with following objectives:

- Physical phenotyping: A high-throughput method for physical characterization of cells
- Biological phenotyping: Characterization of cells to analyze their metastatic potential
- Modeling: Statistical functional methods for physical/biological cell classification
- Predicting the metastatic potential of cells

This work offers a complete mapping and correlation of different characteristics (biological, physical and genetic) of selected cells lines to model their metastatic potential for improved diagnostic and prognostic capabilities. It aims providing the required tool to predict metastatic potential of a cell by means of physical properties; a method that is faster, cheaper and better suited for point-of-care applications.
For data processing, the needed statistical and computational classification and prediction methods require training with large-scale, heterogeneous, continuous (functional) and spatial datasets including high numbers of cells' physical and biological properties. A very limited number of models are available for description, visualization, classification and prediction of quantitative data involving continuous (functional) heterogeneous data. Challenges are, in one hand, to use statistical (including functional classification, regression methods) tools able to compute in a non-costly way correlation among huge amounts of data, for training and accuracy prediction. A main issue when analyzing our massive data is to use statistical tools including functional classification, regression methods) able to compute in a non-costly way correlation among huge amounts of data and the ultimate goal of predicting the metastatic potential of cells by physical characterization of cells.

The work is interdisciplinary (biophysicists, biologists with analysts and statisticians) and is in collaboration with University of Tokyo (Bio-MEMS technology in SMMiL-E ; Seeding Microsystems in Medecine in Lille - European-Japanese Technologies against Cancer) project: http://www.ircl.org/programme-smmil-e/).

References:

[1] Fidler IJ. Nat Rev Cancer, 2003, 3, 453–458.
[2] Punnoose EA, Atwal SK, Spoerke JM, Savage H, Pandita A, et al.,. PLoS One. 2010, 5:e12517.
[3] Lagadec C; Vlashi E; Della Donna L; Dekmezian C; Pajonk F. Stem Cell, 2012, 30(5), 833-44.
[4] Millner L, Linder M, Valdes R. Ann Clin Lab Sci. 2013, 43(3), 295–304.
[5] S. E. Cross, Y.-S. Jin, J. Rao, and J. K. Gimzewski, Nature Nanotech, 2007, 2, 12, 780–783.
[6] J. Guck, S. Schinkinger, B. Lincoln, F. Wottawah, S. Ebert, et al.,, Biophys. J, 2005, 88, 5, 3689–3698.
[7] A. Han, L. Yang, and A. B. Frazier, Clinical Cancer Research, 2007, 13, 1, 139–143.

Speaker: Aurore Delaigle (University of Melbourne)

Title: Classification and clustering of functional data using projection

Abstract: We show that, in the functional data context, by appropriately exploiting the functional nature of the data, it is possible to classify and cluster the observations asymptotically perfectly. We demonstrate that this level of performance can often be achieved as the data are projected on a carefully chosen finite dimensional space.

In the clustering case, we propose an iterative algorithm to choose the projection functions in a way that optimises clustering performance, where, to avoid peculiar solutions, we use a weighted least-squares criterion. We apply our iterative clustering procedure on simulated and real data.

Speaker: Marie-Helene Descary (Universite de Montreal)

Title: Functional Data Analysis by matrix completion

Abstract: Functional data are complex data objects, such as curves and surfaces, that can be seen as realizations of a random function. Covariance operators are at the core of the analysis of such data, since functional PCA is the canonical dimension reduction technique used to go from infinite to finite dimensions. In this talk, we consider the problem of nonparametric estimation of a covariance operator, given a sample of discretely observed functional data, for two different setups. In the first setup, we suppose that the observed data arise as the sum of two uncorrelated components, a smooth one representing the global variations of the data and a rough one representing the local variations of the data, and our focus is to recover the covariance operator of the smooth component. In the second setup, we suppose that the discretely observed functional data are fragments, i.e. that the curves are not observed on the whole domain of definition [0; 1] but only on a subinterval of length strictly smaller than 1, and we want to recover the covariance operator on the whole unit square. For each setup, we show that the estimation problem translates to a low-rank matrix completion problem and construct a nonparametric estimator via rank-constrained least squares. We illustrate our method by simulation and analysis of real data and provide theory to show the validity of the method.

Speaker: Idris Eckley (Lancaster University)

Title: Anomalies, changepoints and exoplanets

Abstract: The detection of changepoints and anomalies are of growing importance for many applications, primarily due to the abundance of sensors within contemporary systems and devices. Such sensors are capable of generating a large amount of data, necessitating computationally efficient methods for their analysis. To date, much of the statistical literature has been concerned with the detection of point anomalies, whilst the problem of detecting anomalous segments – often called collective anomalies – has been relatively neglected. We will introduce work that seeks to address this gap, developing an approach which can differentiate

between both point anomalies and anomalous segments in linear time. We will also explore its utility on the challenging problem of detecting exoplanets using data from the Kepler telescope. If time permits, we will also consider the extension of the approach to higher dimensional settings.

This is joint work with Alex Fisch and Paul Fearnhead

Speaker: Frederic Ferraty (University of Toulouse)

Title: Theoretical foundations for functional local linear regression

Abstract: Local linear regression is one of the most popular regression method when the predictor is a finite-dimensional covariate. It is well known that the local linear regression outperforms the usual kernel estimator and the literature dealing with this topic is very dense. To our knowledge (and surprisingly) there are only two papers extending the local linear regression to the situation when one considers a functional predictor. Problem: the theoretical developments of one of these works is absolutely false where in the second one, the authors require strong assumptions with respect to the distribution of the functional predictor. Even if the infinite-dimensional feature of the predictor makes challenging the asymptotics in the functional local linear regression, it is clear that this topic is still underdeveloped. So this talk aims to bring a more relevant response by proposing new theoretical developments. As a by-product, we also provide the asymptotics for the derivative of the functional linear regression operator.

(Joint work with S. Nagy)

Speaker: Gery Geenens (University of New South Wales)

Title: The Hellinger Correlation; and a Functional Analogue?

Abstract: In this work, the defining properties of a valid measure of the dependence between two random variables are reviewed and complemented with two original ones, shown to be more fundamental than other usual postulates. While other popular choices are proved to violate some of these requirements, a class of dependence measures satisfying all of them is identified. One particular measure, that we call the Hellinger correlation, appears as a natural choice within that class due to both its theoretical and intuitive appeal. A simple and efficient nonparametric estimator for that quantity is proposed. Synthetic and real-data examples finally illustrate the descriptive ability of the measure, which can also be used as test statistic for exact independence testing. At the end of the talk, I will open the discussion about possible extension to defining and measuring dependence between functional random variables.

Speaker: Rob Hyndman (Monash University)

Title: Data visualization for functional time series

Abstract: Any good data analysis begins with a careful graphical exploration of the observed data. For functional time series data, this area of statistical analysis has been largely neglected. I will look at the tools that are available such as rainbow plots and functional box plots, and propose several new tools including functional ACF plots, functional season plots, calendar plots, and embedded pairwise distance plots. These will be illustrated using pedestrian count data in Melbourne, smart metre data from Ireland, and mortality data from France.

Speaker: Wei Huang (University of Melbourne)

Title: Title: Estimating the covariance function from incompletely observed functional data

Abstract: We consider the problem of estimating the covariance function of functional data which are only observed on a subset of their domain, for example in the form of fragments observed on a small interval. Typically, in this setting, no curve is observed on the entire domain so that the empirical covariance function or smooth versions of it can be computed only on a subset of its domain which typically consists in a diagonal band. Estimating nonparametrically the covariance function consistently outside that subset is an extrapolation problem and require identifiablity conditions. We establish such conditions and propose a tensor product series approach for estimating it. We show that our estimator is consistent on the entire domain and illustrate its finite sample properties on simulated and real data.

Speaker: Ci-Ren Jiang (Academia Sinica)

Title: Predicting One-day-ahead Wind Power Capacity Factor via Functional Inverse Regression

Abstract: Fisher's linear discriminant analysis (LDA) is extended to both densely recorded functional data and sparsely observed longitudinal data for general c-category classification problems. An efficient approach is proposed to identify the optimal LDA projections in addition to managing the noninvertibility issue of the covariance operator emerging from this extension. To tackle the challenge of projecting sparse data to the LDA directions, a conditional expectation technique is employed. The asymptotic properties of the proposed estimators are investigated, and asymptotically perfect classification is shown to be achievable in certain circumstances. The performance of this new approach is further demonstrated with both simulated data and real examples.

Speaker: Pavel Krupskiy (University of Melbourne)

Title: Conditional Normal Copulas and Their Use in Applications

Abstract: We propose a new class of copulas which is a generalization of conditional independence models. In these models, dependence among observed variables is modeled using one or several unobserved factors. Conditional on these factors, the distribution of these variables is given by the Gaussian copula. This structure allows one to build flexible and parsimonious models for data with complex dependence structures, such as data with spatial or temporal dependence, or to construct copulas with dynamic dependencies. With different choices of the factors one can obtain tail dependence and/or asymmetry, and parameter estimation is quite fast even in high dimensions when using the maximum likelihood approach. We show some interesting special cases of the proposed class of copulas and illustrate ideas with an empirical study.

Speaker: Dominik Liebl (University of Bonn)

Title: Points of Impact in Generalized Linear Models with Functional Predictors

Abstract: Generalized linear models with function-valued predictor variables and scalar outcomes belong to the well-established and widely used methodological toolbox in functional data analysis. In many applications, however, only specific locations or time-points of the functional predictors have an impact on the outcome. The selection of such points of impact constitutes a particular variable selection problem, since the high correlation in the functional predictors violates the basic assumptions of existing high-dimensional variable selection procedures. In this paper we introduce a generalized linear regression model with functional predictors evaluated at unknown points of impact which need to be estimated from the data alongside the model parameters. We propose a threshold-based and a fully data-driven estimator, establish the identifiability of our model, derive the convergence rates of our point of impact estimators, and develop the asymptotic normality of the estimators of the linear model parameters. The finite sample properties of our estimators are assessed by means of a simulation study. Our methodology is motivated by a psychological case study in which the participants were asked to continuously rate their emotional state while watching an affective online video on the persecution of African albinos. Accompanying supplementary materials are available online.

Joint work with Dominik Poss and Alois Kneip

Speaker: Eardi Lila (University of Cambridge)

Title: Statistics on functional data and covariance operators in linear inverse problems

Abstract: In this talk, we present a framework for the statistical analysis of functional data in a setting where these objects cannot be fully observed, but only indirect and noisy measurements are available, namely an inverse problem setting. The samples can either be unconstrained functional data or objects living in non-Euclidean spaces, such as covariance operators. To illustrate the proposed ideas, we will show an application to medical imaging.

Speaker: Steve Marron (University of North Carolina, Chapel Hill)

Title: Object oriented data analysis

Abstract: The rapid change in computational capabilities has made Big Data a major modern statistical challenge. Less well understood is the rise of Complex Data as a perhaps greater challenge. Object Oriented Data Analysis (OODA) is a framework for addressing this, in particular providing a general approach to the definition, representation, visualization and analysis of Complex Data. The notion of Object Oriented Data Analysis generally guides data analysis, through providing a useful terminology for interdisciplinary discussion of the many choices needed in many modern complex data analyses. OODA ideas are applied in the challenging area of statistical analysis of populations of shapes, motivated by medical image analysis. A particular challenge is automatic segmentation in high noise, low contrast situations. A fundamental component is various approaches to extending the linear method principal component analysis to nonlinear manifolds.

Speaker: Debashis Paul (University of California, Davis)

Title: Modeling subject-specific dynamics through ODEs

Abstract: We consider modeling non-autonomous dynamical systems for a group of subjects. The proposed model involves a common baseline gradient function and a multiplicative time-dependent subject-specific effect that accounts for phase and amplitude variations in the rate of change across subjects. The baseline gradient function is represented in a spline basis and the subject-specific effect is modeled as a polynomial in time with random coefficients. We establish appropriate identifiability conditions and propose an estimator based on the hierarchical likelihood. We prove consistency and asymptotic normality of the proposed estimator under a regime of moderate-to-dense observations per subject. Simulation studies and an application to the Berkeley Growth Data demonstrate the effectiveness of the proposed methodology.

(Joint work with Jie Peng and Siyuan Zhou)

Speaker: Jim Ramsay (McGill University)

Title: The Surprising Landscape beyond Inner Product Spaces

Abstract: We depend heavily on formulating problems within Hilbert spaces where we freely use inner products. This may be the main reason why there are practically no probability distributions other than the Gaussian available for higher dimensional data spaces.

But what if data spaces do not have an origin, and therefore are not amenable to using inner products? Suppose, for example, that the function spaces that we envisage only provide angles and distances and are therefore Euclidean in the strict sense? Or are non-flat manifolds and therefore only locally strictly Euclidean? Some examples are offered, some possible approaches to defining probability distributions over such spaces proposed.

Speaker: Matthew Reimherr (PennState University)

Title: Differential Privacy in Functional Data Analysis

Abstract: In statistical privacy (or statistical disclosure control) the goal is to minimize the potential for identification of individual records or sensitive characteristics while at the same time ensuring that the released information provides accurate and valid statistical inference. Differential Privacy, DP, has emerged as a mathematically rigorous definition of risk and more broadly as a framework for releasing privacy enhanced versions of a statistical summary. This work develops an extensive theory for achieving DP with functional data or function valued parameters more generally. Functional data analysis, FDA, as well as other branches of statistics and machine learning often deal with function valued parameters. Functional data and/or functional parameters may contain unexpectedly large amounts of personally identifying information, and thus developing a privacy framework for these areas is critical in the era of big data. Our theoretical framework is based on densities over function spaces, which is of independent interest to FDA researchers, as densities have proven to be challenging to define and utilize for FDA models. Of particular interest to researchers working in statistical disclosure control, we demonstrate how even small amounts of over smoothing or regularizing can produce releases with substantially improved utility. We carry out extensive simulations to examine the utility of privacy enhanced releases and consider applications to Diffusion Tensor Imaging and high-resolution 3D facial imaging.

Speaker: Damla Senturk (University of California, Los Angeles)

Title: Functional Principal Components Approaches to High-dimensional EEG Data

Abstract: We will outline approaches to modeling Electroencephalography (EEG) data as a high-dimensional and highly-structured functional data and touch up on open problems in the field. More specifically, hybrid principal components analysis (HPCA) will be introduced for region-referenced high-dimensional EEG data from a single experimental task. Our motivating example is a word segmentation paradigm in which typically developing (TD) children and children with Autism Spectrum Disorder (ASD) were exposed to a continuous speech stream. For each subject, continuous EEG signals recorded at each electrode were divided into one-second segments and projected into the frequency domain via Fast Fourier Transform. Following a spectral principal components analysis, the resulting data consist of region-referenced principal power indexed regionally by scalp location and functionally across frequencies and one-second segments. Standard EEG power analyses often collapse information across the segment dimension by averaging power across segments and concentrating on specific frequency bands. We propose a hybrid principal component analysis (HPCA) for region-referenced functional EEG data which utilizes both vector and functional principal components analyses and does not collapse information along any of the dimensions of the data. The proposed decomposition assumes weak separability of the higher-dimensional covariance process and utilizes a product of one dimensional eigenvectors and eigenfunctions, obtained from the regional and functional marginal covariances, to represent the observed data, providing a computationally feasible nonparametric approach. We will also expand the discussion to open problems in applications to EEG data. Possible extensions include analysis of EEG data from multiple experimental tasks, across multiple biobehavioral modalities and over longitudinal visits. We will specifically touch upon joint analysis of two popular biobehavioral modalities in ASD research, EEG and eye-tracking (ET), which are both noninvasive, low cost and widely available.

Speaker: Han Lin Shang (Australian National University)

Title: Forecasting of density functions with an application to cross-sectional and intraday returns

Abstract: Functional time series analysis deals with a time series of functions. A relevant case of functional time series analysis is when the observed functions are density functions. Such data are ubiquitous, as density values are nonnegative and have a constrained integral. Due to the inherent constraints, densities do not live in a vector space but live in a nonnegative-valued constrained space bounded between zero and one. To remedy the problem, we introduce a compositional data analytic approach and view a density function as an example of infinite-dimensional compositional data. Via log-ratio transformation, the densities are mapped into a vector space, where a functional time series forecasting method can be applied to model and forecast density functions. The unconstrained forecast density functions are then mapped back to the constrained space, via inverse log-ratio transformation. Also, we extend the log quantile density transformation of Petersen and Muller (2016) to the case where the densities have

different support and densities converge to zero at the boundaries of the support. Using the monthly cross-sectional and intraday financial data, we demonstrate the finite-sample performance of the proposed methods and compare them with two existing methods, namely the dynamic functional principal component regression of Horta and Ziegelmann (2018) and the skewed generalized t distribution of Wang (2012).

Speaker : Jian Qing Shi (Newcastle University, UK)

Title : Modelling function-valued processes with nonseparable covariance structure

Abstract : Separability of the covariance structure is a common assumption for function-valued processes defined on two- or higher-dimensional domains. This assumption is often made to obtain an interpretable model or due to difficulties in modelling a potentially complex covariance structure, especially in the case of sparse designs. We suggest using Gaussian processes with flexible covariance kernels whose parameters are assumed to follow an underlying continuous function over the input space. This approach enables us to model random process with nonstationary, nonseparable covariance structure with interpretable parameters. We show that the leading eigensurfaces of the estimated covariance function can explain well the main modes of variation in the functional data, including the interactions between the inputs. The results are demonstrated by simulation studies and by an application to human fertility data.

Speaker: Katharine Turner (Australian National University)

Title: Persistent Homology Transform meets Functional Data Analysis

Abstract: The persistent homology transform (PHT) creates a topological summary of a geometric shape. For each unit vector we can consider dot product function over the shape with that vector. Taking sub level sets effectively scans the object with respect to that direction. The persistent homology transform associates to each direction a topological summary called a persistence diagram that describes how the homology of that shape evolves through that sub-level set scan. Persistence diagrams are effectively multisets in the plane with abstract copies of the diagonal. In the PHT these points vary continuously as we vary the scanning direction vector. I would be interested considering the PHT as a set of pairs of functions over subsets of the sphere, where these functions are the coordinates of a point continuously varying over the corresponding persistence diagrams. Potentially we could use techniques for functional data analysis to describe family of shapes. In the talk I will introduce the PHT and present some thoughts about how to relate to functional data analysis, leaving plenty of room for discussion.

Speaker: Fang Yao (University of Toronto/Peking University)

Title: Mixture inner product spaces and application to functional data (by Zhenhua Lin, Hans G. Mueller and **Fang Yao**)

Abstract: We introduce the concept of mixture inner product spaces associated with a given separable Hilbert space, which features 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. 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. 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 an adaptive trajectory-specific dimension that corresponds to a latent random variable and 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.