Functional data analysis

Jane-Ling Wang,[1] Jeng-Min Chiou,[2] and Hans-Georg Müller[1]

[1]Department of Statistics/University of California, Davis, USA, 95616
[2]Institute of Statistical Science/Academia Sinica,Tapei, Taiwan, R.O.C.

Annu. Rev. Statist. 2015. ?:1–41

This article's doi:
10.1146/((please add article doi))

Keywords

Functional principal component analysis, functional correlation, functional linear regression, functional additive model, clustering and classification, time warping

Abstract

With the advance of modern technology, more and more data are being recorded continuously during a time interval or intermittently at several discrete time points. They are both examples of "functional data", which have become a commonly encountered type of data. Functional Data Analysis (FDA) encompasses the statistical methodology for such data. Broadly interpreted, FDA deals with the analysis and theory of data that are in the form of functions. This paper provides an overview of FDA, starting with simple statistical notions such as mean and covariance functions, then covering some core techniques, the most popular of which is Functional Principal Component Analysis (FPCA). FPCA is an important dimension reduction tool and in sparse data situations can be used to impute functional data that are sparsely observed. Other dimension reduction approaches are also discussed. In addition, we review another core technique, functional linear regression, as well as clustering and classification of functional data. Beyond linear and single or multiple index methods we touch upon a few nonlinear approaches that are promising for certain applications. They include additive and other nonlinear functional regression models, and models that feature time warping, manifold learning, and empirical differential equations. The paper concludes with a brief discussion of future directions.

Contents

1. Introduction

Functional data analysis (FDA) deals with the analysis and theory of data that are in the form of functions, images and shapes, or more general objects. The atom of functional data is a function, where for each subject in a random sample one or several functions are recorded. While the term "functional data analysis" was coined by Ramsay (1982) and Ramsay & Dalzell (1991), the history of this area is much older and dates back to Grenander (1950) and Rao (1958). Functional data are intrinsically infinite dimensional. The high intrinsic dimensionality of these data poses challenges both for theory and computation, where these challenges vary with how the functional data were sampled. On the other hand, the high or infinite dimensional structure of the data is a rich source of information, which brings many opportunities for research and data analysis.

First generation functional data typically consist of a random sample of independent real-valued functions, $X_1(t), \ldots, X_n(t)$, on a compact interval $I = [0, T]$ on the real line. Such data have also been termed curve data (Gasser et al. 1984; Rice & Silverman 1991; Gasser & Kneip 1995). These real-valued functions can be viewed as the realizations of a one-dimensional stochastic process, often assumed to be in a Hilbert space, such as $L^2(I)$. Here a stochastic process $X(t)$ is said to be an L^2 process if and only if it satisfies $E(\int_I X^2(t)dt) < \infty$. While it is possible to model functional data with parametric approaches, usually mixed effects nonlinear models, the massive information contained in the infinite dimensional data and the need for a large degree of flexibility, combined with a natural ordering (in time) within a curve datum facilitate non- and semi-parametric approaches, which are the prevailing methods in the literature as well as the focus of this paper. Smoothness of individual random functions (realizations of a stochastic process), such as existence of continuous second derivatives, is often imposed for regularization, and is especially useful if nonparametric smoothing techniques are employed, as is prevalent in functional data analysis

In this paper, we focus on first generation functional data with brief a discussion of next generation functional data in Section 6. Here next generation functional data refers to functional data that are part of complex data objects, and possibly are multivariate, correlated, or involve images or shapes. Examples of next generation functional data include brain and neuroimaging data. A separate entry on functional data approaches for neuroimaging data is available in this issue of the Annual Reviews (Link to John Aston's contribution). For a brief discussion of next generation functional data, see page 23 of a report (http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf) of the London workshop on the Future of Statistical Sciences held in November 2013.

Although scientific interest is in the underlying stochastic process and its properties, in reality this process is often latent and cannot be observed directly, as data can only be collected discretely over time, either on a fixed or random time grid. The time grid can be dense, sparse, or neither; and may vary from subject to subject. Originally, functional data were regarded as samples of fully observed trajectories. A slightly more general assumption is that functional data are recorded on the same dense time grid of ordered times $t_1, \ldots, t_p$ for all n subjects. If the recording is done by an instrument, such as an EEG or fMRI recording device, the time grid is usually equally spaced, that is $t_j - t_{j-1} = t_{j+1} - t_j$ for all j. In asymptotic analysis, the spacing $t_{j+1} - t_j$ is assumed to approach zero as n tends to infinity, hence $p = p_n$ is a sequence that tends to infinity. While large p leads to a high-dimensional problem, it also means more data are available, so here this is a blessing rather than a curse. This blessing is realized by imposing a smoothness assumption on the L^2 processes, so that information from measurements at neighboring time points can be pooled to overcome the curse of dimensionality. Thus, smoothing serves as a tool for regularization.

While there is no formal definition of "dense" functional data, the convention has been to declare functional data as densely (as opposed to sparsely) sampled when p_n converges to infinity fast enough to allow the corresponding estimate for the mean function $\mu(t) = EX(t)$, where X is the underlying process, to attain the parametric $\sqrt{n}$ convergence rate for standard metrics, such as the L^2 norm. Sparse functional data arise in longitudinal studies where subjects are measured at different time points and the number of measurements n_i for subject i may be bounded away from infinity, i.e., $\sup_{1 \leq i \leq n} n_i < C < \infty$ for some constant C. A rigorous definition of the types of functional data based on their sampling plans is still lacking. See Zhang & Wang (2014) for a possible approach, with further details in Section 2 below.

In reality, many observed data are contaminated by random noise, referred to as measurement errors, which are often assumed to be independent across and within subjects. Measurement errors can be viewed as random fluctuations around a smooth trajectory, or as actual errors in the measurements. A strength of FDA is that it accommodates measurement errors easily, as for each subject one observes repeated measurements. An interesting, but not surprising, phenomenon in FDA is that the methodology and theory, such as convergence rates, varies with the measurement schedule (sampling plan).

Intriguingly, sparse and irregularly sampled functional data that we synonymously refer to as longitudinal data, such as the CD4 count data for which a Spaghetti plot is shown in Figure 1 for 25 subjects, typically require more effort in theory and methodology as compared to densely sampled functional data, such as the traffic data in Figure 5, which are recorded continuously. For the CD4 count data, a total of 369 subjects were included with the number of measurements ranging from 1 to 12, with median (mean) number of measurements 6 (6.44). This is an example of sparse functional data measured at an irregular and different time schedule for each individual. Functional data that are observed continuously without errors are the easiest type to handle as theory for stochastic processes,

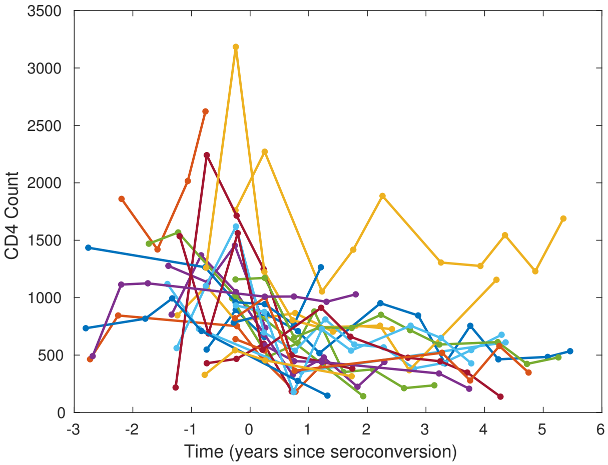

Figure 1

Spaghetti plot for sparsely recorded CD4 count data for 25 subjects

such as functional laws of large numbers and functional central limit theorems, are readily applicable. A comparison of the various approaches will be presented in Section 2.

One of the challenges in functional data analysis is the inverse nature of functional regression and most functional correlation measures. This is triggered by the compactness of the covariance operator, which leads to unbounded inverse operators. This challenge will be discussed further in Section 3, where extensions of classical linear and generalized linear models to functional linear and generalized functional linear models will be reviewed. Since functional data are intrinsically infinite dimensional, dimension reduction is key for data modeling and analysis. The principal component approach will be explored in Section 2 while several approaches for dimension reduction in functional regression will be discussed in Section 3.

Clustering and classification of functional data are useful and important tools in FDA with wide ranging applications. Methods include extensions of classical k-means and hierarchical clustering, Bayesian and model-based approaches to clustering, as well as classification via functional regression based and functional discriminant analysis. These topics will be explored in Section 4. The classical methods for functional data analysis have been predominantly linear, such as functional principal components or the functional linear model. As more and more functional data are being generated, it has emerged that many such data have inherent nonlinear features that make linear methods less effective. Sections 5 reviews some nonlinear approaches to FDA, including time warping, non-linear manifold modeling, and nonlinear differential equations to model the underlying empirical dynamics.

A well-known and well-studied nonlinear effect is time warping, where in addition to the common amplitude variation one also considers time variation. This creates a basic non-identifiability problem. Section 5.1 will provide a discussion of these foundational issues. A more general approach to model nonlinearity in functional data that extends beyond time warping and includes many other nonlinear features that may be present in longitudinal data is to assume that the functional data lie on a nonlinear (Hilbert) manifold. The starting point for such models is the choice of a suitable distance and ISOMAP (Tenenbaum, De Silva & Langford 2000) or related methods can then be employed to uncover the manifold structure and define functional manifold means and components. These approaches will be described in Section 5.2. Modeling of time-dynamic systems with differential equations that are learned from many realizations of the trajectories of the underlying stochastic process and the learning of nonlinear empirical dynamics such as dynamic regression to the mean or dynamic explosivity is briefly reviewed in Section 5.3. Section 6 concludes this review with a brief outlook on the future of functional data analysis.

Research tools that are useful for handing functional data include various smoothing methods, notably kernel, local least squares and spline smoothing for which various excellent reference books exist (Wand & Jones 1995; Fan & Gijbels 1996; Eubank 1999; de Boor 2001), functional analysis (Conway 1994; Hsing & Eubank 2015) and stochastic processes (Ash & Gardner 1975). Several software packages are publicly available to analyze functional data, including software at the Functional Data Analysis website of James Ramsay (http://www.psych.mcgill.ca/misc/fda/), the `fda` package on the CRAN project of `R` (R Core Team 2013) (http://cran.r-project.org/web/packages/fda/fda.pdf), the Matlab package `PACE` on the website of the Statistics Department of the University of California, Davis (http://www.stat.ucdavis.edu/PACE/), and the `R` package `refund` on functional regression (http://cran.r-project.org/web/packages/refund/index.html).

This review is based on a subjective selection of topics in FDA that the authors have worked on or find of particular interest. We do not attempt to provide an objective or comprehensive review of this fast moving field and apologize in advance for any omissions of relevant work. Interested readers can explore the various aspects of this field through several monographs (Bosq 2000; Ramsay & Silverman 2005; Ferraty & Vieu 2006; Wu & Zhang 2006; Ramsay, Hooker & Graves 2009; Horvath & Kokoszka 2012; Hsing & Eubank 2015) and review articles (Rice 2004; Zhao, Marron & Wells 2004; Müller 2005, 2008; Ferraty & Vieu 2006). Several special journal issues were devoted to FDA including a 2004 issue of *Statistica Sinica* (issue 3), a 2007 issue in *Computational Statistics and Data Analysis* (issue 3), and a 2010 issue in *Journal of Multivariate analysis* (issue 2).

2. Mean and Covariance Function, and Functional Principal Component Analysis

In this section, we focus on first generation functional data that are realizations of a stochastic process $X(\cdot)$ that is in L^2 and defined on the interval I with mean function $\mu(t) = E(X(t))$ and covariance function $\Sigma(s,t) = \text{cov}(X(s), X(t))$. The functional framework can also be extended to L^2 processes with multivariate arguments. The realization of the process for the ith subject is $X_i = X_i(\cdot)$, and the sample consists of n independent subjects. For generality, we allow the sampling schedules to vary across subjects and denote the sampling schedule for subject i as $t_{i1}, \ldots, t_{in_i}$ and the corresponding observations as $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})$, where $X_{ij} = X_i(t_{ij})$. In addition, we allow the measurement of X_{ij} to be contaminated by random noise e_{ij} with $E(e_{ij}) = 0$ and $\text{var}(e_{ij}) = \sigma_{ij}^2$, so the actual observed value is $Y_{ij} = X_{ij} + e_{ij}$, where e_{ij} are independent across i and j and often termed "measurement errors".

It is often assumed that the errors are homoscedastic with $\sigma_{ij}^2 = \sigma^2$, but this is is not strictly necessary, as long as $\sigma_{ij}^2 = \text{var}(e(t_{ij}))$ can be regarded as the discretization of a smooth variance function $\sigma^2(t)$. We observe that measurement errors are realized only at those time points t_{ij} where measurements are being taken. Hence these errors do not form a stochastic process $e(t)$ but rather should be treated as discretized data e_{ij}. However, in order to estimate the variance σ_{ij}^2 of e_{ij} it is often convenient to assume that there is a latent smooth function $\sigma(t)$ such that $\sigma_{ij} = \sigma^2(t_{ij})$.

Estimation of Mean and Covariance Functions. When subjects are sampled at the same time schedule, i.e., $t_{ij} = t_j$ and $n_i = m$ for all i, the observed data are m-dimensional multivariate data, so the mean and covariance can be estimated empirically at the measurement times by the sample mean and sample covariance, $\hat{\mu}(t_j) = \frac{1}{n}\sum_{i=1}^n Y_{ij}$, and $\hat{\Sigma}(t_k, t_l) = \frac{1}{n}\sum_{i=1}^n (Y_{ik} - \hat{\mu}(t_{ik}))(Y_{il} - \hat{\mu}(t_{il}))$, for $k \neq l$. Data that are missing completely at random (for further details on missingness see Little & Rubin 2014) can be handled easily by adjusting the available sample size at each time point t_j for the mean estimate or by adjusting the sample sizes of available pairs at (t_k, t_l) for the covariance estimate. An estimate of the mean and covariance functions on the entire interval I can then be obtained by smooth interpolation of the corresponding sample estimates or by mildly smoothing over the grid points. Once we have a smoothed estimate $\hat{\Sigma}$ of the covariance function Σ, the variance of the measurement error at time t_j can be estimated as $\hat{\sigma}^2(t_j) = \frac{1}{n}\sum_{i=1}^n (Y_{ij} - \hat{\mu}(t_j))^2 - \hat{\Sigma}(t_j, t_j)$, because $\mathrm{var}(Y(t)) = \mathrm{var}(X(t)) + \sigma^2(t)$.

When the sampling schedule of subjects differs, the above sample estimates cannot be obtained. However, one can borrow information from neighboring data and across all subjects to estimate the mean function, provided the sampling design combining all subjects, i.e. $\{t_{ij} : 1 \leq i \leq n,\ 1 \leq j \leq n_i\}$, is a dense subset of the interval I. Then a nonparametric smoother, such as a local polynomial estimate (Fan & Gijbels 1996), can be applied to the scatter plot $\{(t_{ij}, Y_{ij}) : i = 1, \ldots, n,$ and $j = 1, \ldots, n_i\}$ to smooth Y_{ij} over time; this will yield consistent estimates of $\mu(t)$ for all t. Figure 2 shows the scatter plot of the pooled CD4 counts for all 369 AIDS patients, together with the estimated mean function based a local linear smoother with bandwidth 0.3 year. The shape of the mean function reveals that CD4 counts were stable around 1,000 six months before seroconversion (time 0) but decline sharply six months before and after seroconversion, and then stabilize again after one year of seroconversion.

Likewise, the covariance can be estimated on $I \times I$ by a two-dimensional scatter plot smoother $\{(t_{ik}, t_{il}), u_{ikl} : i = 1, \ldots, n;\ k, l = 1, \ldots, n_i, k \neq l\}$ to smooth u_{ikl} against (t_{ik}, t_{il}), where $u_{ikl} = (Y_{ik} - \hat{\mu}(t_{ik}))(Y_{il} - \hat{\mu}(t_{il}))$ are the "raw" covariances. We note that the diagonal raw covariances where $k = l$ are removed from the 2D scatter plot prior to the smoothing step because these include an additional term that is due to the variance of the measurement errors in the observed Y_{ij}. Indeed, once an estimate $\hat{\Sigma}$ for Σ is obtained, the variance $\sigma^2(t)$ of the measurement errors can be obtained by smoothing $Y_{ij} - \hat{\mu}(t_{ij})^2 - \hat{\Sigma}(t_{ij})$ against t_{ij} across time. Figure 3 displays the scatter plot of the raw covariances and the smoothed estimate of the covariance surface $\Sigma(\cdot, \cdot)$ using a local linear bivariate smoother with bandwidth of 1.4 years together with the smoothed estimate of $\mathrm{var}(Y(t))$. The estimated variance of $\sigma^2(t)$ is the distance between the estimated $\mathrm{var}(Y(t))$ and the estimated covariance surface. Another estimate for σ^2 under the homoscedasticity assumption is discussed in Yao, Müller & Wang (2005a).

The above smoothing approach is based on a scatter plot smoother which assigns equal weights to each observation, therefore subjects with a larger number of repeated observations receive more total weight, and hence contribute more toward the estimates of the mean and covariance functions. An alternative approach employed in Li & Hsing (2010) is to assign equal weights to each subject. Both approaches are sensible. A question is

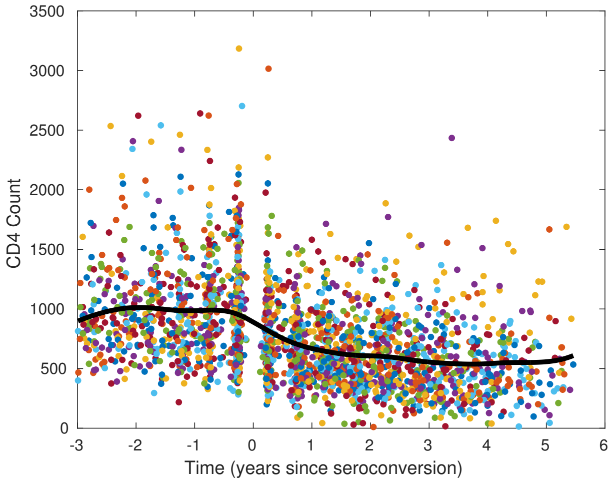

Figure 2

Pooled CD4 count data for 369 subjects and estimated mean function

which one would be preferred for a particular design and whether there is a unified way to deal with these two methods and their theory. These issues were recently explored in a manuscript (Zhang & Wang 2014), employing a general weight function and providing a comprehensive analysis of the asymptotic properties on a unified platform for three types of asymptotics, L^2 and L^∞ (uniform) convergence as well as asymptotic normality of the general weighted estimates. Functional data sampling designs are further partitioned into three categories, non-dense (designs where one cannot attain the $\sqrt{n}$ rate), dense (where one can attain the $\sqrt{n}$ rate but with a non-neglible asymptotic bias), and ultra-dense (where one can attain the $\sqrt{n}$ rate without asymptotic bias). Sparse sampling scenarios where n_i is uniformly bounded by a finite constant are a special case of non-dense data and lead to the slowest convergence rates. These designs are also referred to as longitudinal designs. The differences in the convergence rates also have ramifications for the construction of simultaneous confidence bands. For ultra dense or some dense functional data, the weighing scheme that assigns equal weights to subjects is generally more efficient than the scheme that assigns equal weight per observation, but the opposite holds for many other sampling plans, including sparse functional data.

Hypothesis Testing and Simultaneous Confidence Bands for Mean and Covariance Functions. Hypothesis testing for the comparison of mean functions μ is of obvious interest. Fan & Lin (1998) proposed a two-sample test and ANOVA test for the mean functions, with further work by Cuevas, Febrero & Fraiman (2004) and Zhang (2013). Other two sample tests have been proposed for distributions of functional data (Hall & Van Keilegom 2007) and for covariance functions (Panaretos, Kraus & Maddocks 2010; Boente, Rodriguez & Sued 2011).

Another inference problem that has been explored is the construction of simultaneous confidence bands for dense (Degras 2008, 2011; Wang & Yang 2009; Cao, Yang & Todem 2012) and sparse (Ma, Yang & Carroll 2012) functional data. However, the problem has

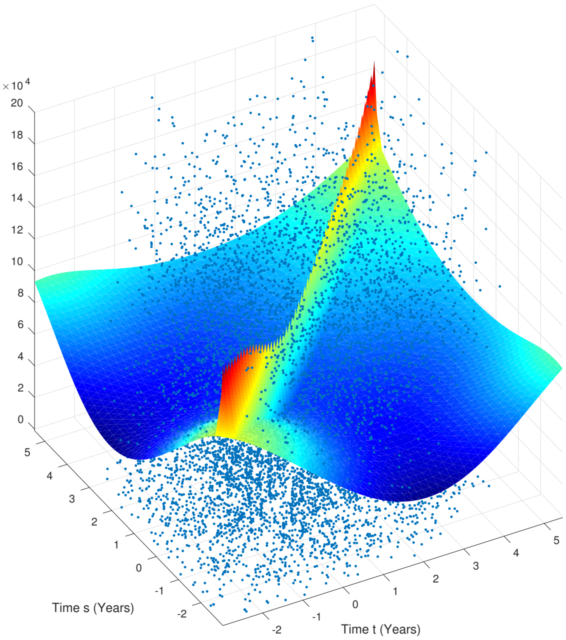

Figure 3

Raw covariances (dots) and fitted smooth covariance surface, obtained by omitting the data on the diagonal, where the diagonal forms a ridge due to the measurement errors in the data. The variance of the measurement error $\sigma^2(t)$ at time t is the vertical distance between the top of the diagonal ridge and the smoothed covariance surface at time t.

not been completely resolved for functional data, due to two main obstacles: The infinite dimensionality of the data and the nonparametric nature of the target function. For the mean function μ, an interesting "phase transition" phenomenon emerges: For ultra-dense data the estimated mean process $\sqrt{n}(\hat{\mu}(t)-\mu(t))$ converges to a mean zero Gaussian process $W(t)$, for $t \in I$, so standard continuous mapping leads to a construction of a simultaneous confidence band based on the distribution of $\sup_t W(t)$. When the functional data are dense but not ultra dense, the process $\sqrt{n}(\hat{\mu}(t) - \mu(t))$ can still converge to a Gaussian process $W(t)$ with a proper choice of smoothing parameter but W is no longer centered at zero due to the existence of asymptotic bias as discussed in Section 1.

This resembles the classical situation of estimating a regression function, say $m(t)$, based on independent scalar response data, where there is a trade off between the bias and variance, so that optimally smoothed estimates of the regression function will have an asymptotic bias. The conventional approach to construct a pointwise confidence interval is based on the distribution of $r_n(\hat{m}(t) - E(\hat{m}(t)))$, where $\hat{m}(t)$ is an estimate of $m(t)$ that converges at the optimal rate r_n. This means that the asymptotic confidence interval derived from it is targeting $E(\hat{m}(t))$ rather than the true target $m(t)$ and therefore is not really viable for inference for $m(t)$.

In summary, the construction of simultaneous confidence bands for functional data requires different methods for ultra-dense, dense, and sparse functional data, where in the latter case one does not have tightness and the rescaling approach of Bickel & Rosenblatt (1973) may be applied. The divide between the various sampling designs is perhaps not unexpected since ultra dense functional data essentially follow the paradigm of parametric inference, where the $\sqrt{n}$ rate of convergence is attained with no asymptotic bias, while dense functional data attains the parametric rate of $\sqrt{n}$ convergence albeit with an asymptotic bias, which leads to challenges even in the construction of pointwise confidence intervals. Unless the bias is estimated separately, removed from the limiting distribution, and proper asymptotic theory is established, which usually requires regularity conditions for which the estimators are not efficient, the resulting confidence intervals need to be taken with a grain of salt. This issue is specific to the bias-variance trade off that is inherited from nonparametric smoothing. Sparse functional data follow a very different paradigm as they allow no more than nonparametric convergence rates, which are slower than $\sqrt{n}$, and the rates depend on the design of the measurement schedule and properties of mean and covariance function as well as the smoother (Zhang & Wang 2014). The phenomenon of nonparametric versus parametric convergence rates as designs get more regular and denser have been characterized as a "phase transition" (Hall, Müller & Wang 2006; Cai & Yuan 2011).

Functional Principal Component Analysis (FPCA). Principal component analysis (Jolliffe 2002) is a key dimension reduction tool for multivariate data that has been extended to functional data and termed functional principal component analysis (FPCA). Although the basic ideas were conceived in Grenander (1950); Karhunen (1946); Loève (1946) and Rao (1958), a more comprehensive framework for statistical inference for FPCA was first developed in a joint Ph.D. thesis of Dauxois and Pousse (1976) at the University of Toulouse (Dauxois, Pousse & Romain 1982). Since then, this approach has taken off to become the most prevalent tool in FDA. This is partly because FPCA facilitates the conversion of inherently infinite-dimensional functional data to a finite-dimensional vector of random scores. Under mild assumptions, the underlying stochastic process can be expressed as a countable sequence of uncorrelated random variables, the functional principal components (FPCs) or scores, which in many practical applications are truncated to a finite vector. Then the tools of multivariate data analysis can be readily applied to the resulting random vector of scores, thus accomplishing the goal of dimension reduction.

Specifically, dimension reduction is achieved through an expansion of the underlying but often not fully observed random trajectories $X_i(t)$ in a functional basis that consists of the eigenfunctions of the (auto)-covariance operator of the process X. With a slight abuse of notation we define the covariance operator as $\Sigma(g) = \int_I \Sigma(s,t)g(s)ds$, for any function $g \in L^2$, using the same notation for the covariance operator and covariance function. Because of the integral form, the (linear) covariance operator is a trace class and hence compact Hilbert-Schmidt operator (Conway 1994). It has real-valued nonnegative eigenvalues λ_j, because it is symmetric and non-negative definite. Under mild assumptions, Mercer's theorem implies that the spectral decomposition of Σ leads to $\Sigma(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$, where the convergence holds uniformly for s and $t \in I$, λ_k are the eigenvalues in descending order and ϕ_k the corresponding orthogonal eigenfunctions. Karhunen and Loève (Karhunen 1946;

Loève 1946) independently discovered the FPCA expansion

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} A_{ik}\phi_k(t), \tag{1}$$

where $A_{ik} = \int_I (X_i(t) - \mu(t))\phi_k(t)dt$ are the functional principal components (FPCs) of X_i, sometimes referred to as *scores*. The A_{ik} are independent across i for a sample of independent trajectories and are uncorrelated across k with $E(A_{ik}) = 0$ and $\mathrm{var}(A_{ik}) = \lambda_k$. The convergence of the sum in (1) holds uniformly in the sense that $\sup_{t\in I} E[X_i(t) - \mu(t) - \sum_{k=1}^{K} A_{ik}\phi_k(t)]^2 \to 0$ as $K \to \infty$. Expansion (1) facilitates dimension reduction as the first K terms for large enough K provide a good approximation to the infinite sum and therefore for X_i, so that the information contained in X_i is essentially contained in the K-dimensional vector $\mathbf{A}_i = (A_{i1}, \dots, A_{iK})$ and one works with the approximated processes

$$X_{iK}(t) = \mu(t) + \sum_{k=1}^{K} A_{ik}\phi_k(t). \tag{2}$$

Analogous dimension reduction can be achieved by expanding the functional data into other function bases, such as spline, Fourier, or wavelet bases. What distinguishes FPCA is that among all basis expansions that use K components for a fixed K, the FPC expansion explains most of the variation in X in the L^2 sense. When choosing K in an estimation setting, there is a trade off between bias (which gets smaller as K increases, due to the smaller approximation error) and variance (which increases with K as more components must be estimated, adding random error). So a model selection procedure is needed, where typically $K = K_n$ is considered to be a function of sample size n and K_n must tend to infinity to obtain consistency of the representation. This feature distinguishes the theory of FPCA from standard multivariate analysis theory.

The estimation of the eigencomponents (eigenfunctions and eigenvalues) in the FPCA framework is straightforward, once mean and covariance of the functional data have been estimated. To obtain the spectral decomposition of the covariance operator, which yields the eigencomponents, one simply approximates the estimated auto-covariance surface $\mathrm{cov}(X(s), X(t))$ on a grid of time points, thus reducing the problem to the corresponding matrix spectral decomposition. The convergence of the estimated eigen-components is obtained by combining results on the convergence of the covariance estimates that are achieved under regularity conditions with perturbation theory (see Chapter VIII of Kato (1980)).

For situations where the covariance surface cannot be estimated at the $\sqrt{n}$ rate, the convergence of the estimated eigen-components is typically influenced by the smoothing method that is employed. Consider the sparse case, where the convergence rate of the covariance surface corresponds to the optimal rate at which a smooth two-dimensional surface can be estimated. Intuition suggests that the eigenfunction, which is a one-dimensional function, should be estimable at the one-dimensional optimal rate for smoothing methods. An affirmative answer is provided in Hall, Müller & Wang (2006), where eigenfunction estimates were shown to attain the better (one-dimensional) rate of convergence, if one is undersmoothing the covariance surface estimate. This phenomenon resembles a scenario encountered in semiparametric inference, e.g. for a partially linear model (Heckman 1986), where a $\sqrt{n}$ rate is attainable for the parametric component if one undersmooths the nonpararmetric component before estimating the parametric component. This undersmoothing can be avoided so that the same smoothing parameter can be employed for both the parametric and nonparametric component if a profile approach (Speckman 1988) is employed to estimate the parametric component. An interesting and still open question is how to construct such a profile approach so that the eigenfunction is the direct target of the estimation procedure, bypassing the estimation of the covariance function.

Another open question is the optimal choice of the number of components K needed for the approximation (2) of the full Karhunen-Loève expansion (1), which gives the best trade-off between bias and variance. There are several ad hoc procedures that are routinely applied in multivariate PCA, such as the scree plot or the fraction of variance explained by the first few PC components, which can be directly extended to the functional setting. Other approaches are pseudo-versions of AIC (Akaike information criterion) and BIC (Bayesian information criterion) (Yao, Müller & Wang 2005a), where typically in practice the latter selects fewer components. Cross-validation with one-curve-leave-out has also been investitgated (Rice & Silverman 1991), but tends to overfit functional data by selecting too large K in (2). A third open question is the optimal choice of the tuning parameters for the smoothing steps in the context of FDA.

FPCA for fully observed functional data was studied in Dauxois, Pousse & Romain (1982), Besse & Ramsay (1986); Silverman (1996), Bosq (2000); Boente & Fraiman (2000); Hall & Hosseini-Nasab (2006), and it was explored for densely observed functional data in Castro, Lawton & Sylvestre (1986); Rice & Silverman (1991); Pezzulli & Silverman (1993) and Cardot (2000). For the much more difficult but commonly encountered situation of sparse functional data, the FPCA approach was investigated in Shi, Weiss & Taylor (1996); Staniswalis & Lee (1998); James, Hastie & Sugar (2000); Rice & Wu (2001); Yao, Müller & Wang (2005a); Yao & Lee (2006); and Paul & Peng (2009). The FPCA approach has also been extended to incorporate covariates (Chiou, Müller & Wang 2003; Cardot 2007; Chiou & Müller 2009) for vector covariates and dense functional data, and also for sparse functional data with vector or functional covariates (Jiang & Wang 2010, 2011) and also to the case of functions in reproducing kernel Hilbert spaces (Amini & Wainwright 2012).

The aforementioned FPCA methods are not robust against outliers because principal component analysis involves second order moments. Outliers for functional data have many different facets due to the high dimensionality of these data. They can appear as outlying measurements at a single or several time points, or as an outlying shape of an entire function. Current approaches to deal with outliers and contamination and more generally visual exploration of functional data include exploratory box plots (Hyndman & Shang 2010; Sun & Genton 2011) and robust versions of FPCA (Crambes, Delsol & Laksaci 2008; Gervini 2008; Bali et al. 2011; Kraus & Panaretos 2012; Boente & Salibián-Barrera 2014). Due to the practical importance of this topic, more research on outlier detection and robust FDA approaches is needed.

Applications of FPCA. The FPCA approach motivates the concept of modes of variation for functional data (Jones & Rice 1992), a most useful tool to visualize and describe the variation in the functional data that is contributed by each eigenfunction. The $k-$th mode of variation is the set of functions

$$\mu(t) \pm \alpha\sqrt{\lambda_k}\phi_k(t),\ t \in I,\ \alpha \in [-A, A],$$

that are viewed simultaneously over the range of α, usually for $A = 2$, substituting estimates for the unknown quantities. Often the eigencomponents and associated modes of variation have compelling and sometimes striking interpretations, such as for the evolution of functional traits (Kirkpatrick & Heckman 1989) and in many other applications (Kneip

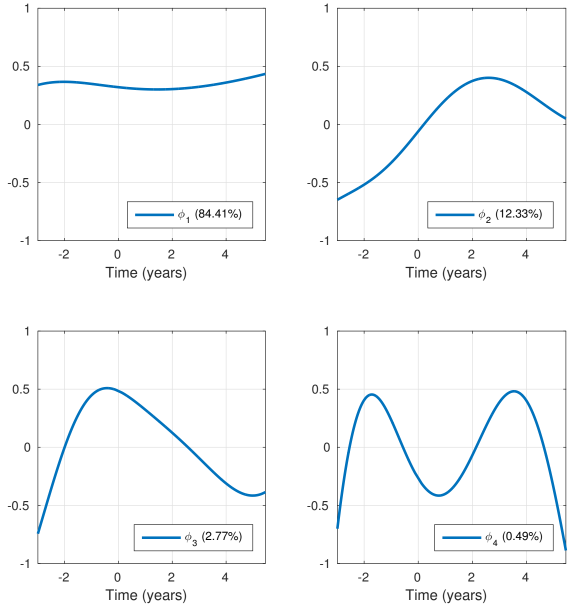

Figure 4

First four eigenfunctions for CD4 data

& Utikal 2001; Ramsay & Silverman 2002). In Figure 4 we provide the first four estimated eigenfunctions for the CD4 counts data.

The first eigenfunction explains 84.41% of the total variation of the data and the second one an additional 12.33% of the data. The remaining two eigenfunctions account for less than 4% of the total variation and are not important. Here the first eigenfunction is nearly constant in time implying that the largest variation between subjects is in the subject specific average magnitude of the CD4 counts, so the random intercept captures the largest variation of the data. The second eigenfunction shows a variation around a piecewise linear time trend with a break point near 2.5 year after seroconversion reflecting that the next largest variation between subjects is a scale difference between subjects along the direction of this piecewise linear function.

FPCA also facilitates functional principal component regression by projecting functional predictors to their first few principal components, then employing regression models with vector predictors. Since FPCA is an essential dimension reduction tool, it is also useful for classification and clustering of functional data (see Section 4).

Last but not least, FPCA facilitates the construction of parametric models that will be more parsimonious. For instance, if the first two principal components explain over 90% of the variation of the data then one can approximate the original functional data with only two terms in the Karhunen-Loeve expansion (1). This can be illustrated with the CD4 counts data, for which a parametric mixed effects model with a piecewise linear time trend (constant before -0.5 year and after 1 year of seroconversion, and linear decline in between) for the fixed effects and a random intercept may suffice to capture the major trend of the data. If more precision is preferred, one could include a second random effect for the piecewise linear basis function with a breakpoint at 2.5 years. This underscores the advantages to use a nonparametric approach such as FDA prior to a model-based longitudinal data analysis for data exploration. The exploratory analysis then may suggest viable parametric models that are more parsimonious than FPCA.

3. Correlation and Regression: Inverse Problems and Dimension Reduction for Functional Data

As mentioned in Section 1, a major challenge in FDA is the inverse problem, which stems from the compactness of the covariance operator Σ that was defined in the previous section, $\Sigma(g) = \int_I \Sigma(s,t)g(s)ds$, for any function $g \in L^2$. If there are infinitely many nonzero, hence positive eigenvalues, then the covariance operator is a one-to one function and the inverse operator of Σ can be determined but it is an unbounded operator and the range space of the covariance operator is a compact set in L^2. This creates a problem to define a bijection, as the inverse of Σ is not defined on the entire L^2 space. Therefore regularization is routinely adopted, for any procedure that involves the inverse on a compact operator. Examples where inverse operators are central include regression and correlation measures for functional data, as Σ^{-1} appears in these methods. This inverse problem was for example addressed for functional canonical correlation in He, Müller & Wang (2000) and He, Müller & Wang (2003), where a solution was discussed under certain constraints on the decay rate of the eigenvalues and the cross covariance operator.

3.1. Functional Correlation

Different functional correlation measures have been discussed in the literature. Functional Canonical Correlation Analysis serves here to demonstrate some of the problems that one encounters in FDA as a consequence of the non-invertibility of compact operators.

Functional Canonical Correlation Analysis (FCCA). Let (X, Y) be a pair of random functions in $L^2(I_X)$ and $L^2(I_Y)$ respectively. The first functional canonical correlation coefficient ρ_1 and its associated weight functions (u_1, v_1) are defined as follows, using the notation $\langle f_1, f_2\rangle = \int_I f_1(t)f_2(t)dt$ for any $f_1, f_2 \in L^2(I)$,

$$\rho_1 = \sup_{u \in L^2(I_X), v \in L^2(I_Y)} \operatorname{cov}(\langle u, X\rangle, \langle v, Y\rangle) = \operatorname{cov}(\langle u_1, X\rangle, \langle v_1, Y\rangle), \tag{3}$$

subject to $\operatorname{var}(\langle u, X\rangle) = 1$ and $\operatorname{var}(\langle v, Y\rangle) = 1$. Analogously for the kth, $k > 1$, canonical correlation ρ_k and its associated weight functions (u_k, v_k),

$$\rho_k = \sup_{u \in L^2(I_X), v \in L^2(I_Y)} \operatorname{cov}(\langle u, X\rangle, \langle v, Y\rangle) = \operatorname{cov}(\langle u_k, X\rangle, \langle v_k, Y\rangle), \tag{4}$$

subject to $\operatorname{var}(\langle u, X\rangle) = 1$, $\operatorname{var}(\langle v, Y\rangle) = 1$, and that the pair $(U_k, V_k) = (\langle u_k, X\rangle, \langle v_k, Y\rangle)$ is uncorrelated to all previous pairs $(U_j, V_j) = (\langle u_j, X\rangle, \langle v_j, Y\rangle)$, for $j = 1, \ldots, k-1$.

Thus, FCCA aims at finding projections in directions u_k of X and v_k of Y such that their linear combinations (inner products) U_k and V_k are maximally correlated, resulting in the series of functional canonical components $(\rho_k, u_k, v_k, U_k, V_k)$, $k \geq 1$, directly extending canonical correlations for multivariate data. Because of the flexibility in the direction

u_1, which is infinite dimensional, overfitting may occur if the number of sample curves is not large enough. Formally, this is due to the fact that FCCA is an ill-posed problem. Introducing the cross-covariance operator $\Sigma_{XY} : L^2(I_Y) \to L^2(I_X)$,

$$\Sigma_{XY} v(t) = \int \text{cov } (X(t), Y(s)) v(s) ds, \tag{5}$$

for $v \in L^2(I_Y)$ and analogously the covariance operators for X, Σ_{XX}, for Y, Σ_{YY}, and using $\text{cov}(\langle u, X\rangle, \langle v, Y\rangle) = \langle u, \Sigma_{XY} Y\rangle$, the kth canonical component in (4) can be expressed as

$$\rho_k = \sup_{u \in L^2(I_X), \langle u, \Sigma_{XX} u\rangle = 1, v \in L^2(I_Y), \langle v, \Sigma_{YY} v\rangle = 1} \langle u, \Sigma_{XY} v\rangle = \langle u_k, \Sigma_{XY} v_k\rangle. \tag{6}$$

Then (6) is equivalent to an eigenanalysis of the operator $R = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$. Existence of the canonical components is guaranteed if the operator R is compact. However, the inverse of a covariance operator and the inverses of $\Sigma_{XX}^{1/2}$ or $\Sigma_{YY}^{1/2}$ are not bounded since a covariance operator is compact under the assumption that the covariance function is square integrable. A possible approach (He, Müller & Wang 2003) is to restrict the domain of the inverse to the range A_X of $\Sigma_{XX}^{1/2}$ so that the inverse of $\Sigma_{XX}^{1/2}$ can be defined on A_X and is a bijective mapping A_X to B_X, under some conditions (e.g., Conditions 4.1 and 4.5 in He, Müller & Wang (2003)) on the decay rates of the eigenvalues of Σ_{XX} and Σ_{YY} and the cross-covariance. Under those assumptions the canonical correlations and weight functions are well defined and exist.

An alternative way to get around the above ill-posed problem is to restrict the maximization in (3) and (4) to discrete l^2 spaces that are restricted to a reproducing kernel Hilbert space instead of working within the entire L^2 space (Eubank & Hsing 2008). In addition to theoretical challenges to overcome the inverse problem, FCCA requires regularization in practical implementations, as only finitely many measurements are available for each subject. If left unregularized, the first canonical correlation will always be one. Unfortunately, the canonical correlations are highly sensitive to the regularization parameter and the first canonical correlation often tends to be too large as there is too much freedom to choose the weights u and v. This makes it difficult to interpret the meaning of the first canonical correlation. The overfitting problem can also be viewed as a consequence of the high-dimensionality of the weight function and was already illustrated in Leurgans, Moyeed & Silverman (1993), who were the first to explore penalized functional canonical correlation analysis. Despite the challenge with overfitting, FCCA can be employed to implement functional regression by using the canonical weight functions u_k, and v_k as bases to expand the regression coefficient function (He, Müller & Wang 2000; He et al. 2010).

Another difficulty with the versions of FCCA proposed so far is that it requires densely recorded functional data so the inner products in (4) can be evaluated with high accuracy. Although it is possible to impute sparsely observed functional data using the Karhunen-Loève expansion (1) before applying any of the canonical correlations, these imputations are not consistent and this leads to a biased correlation estimation. This bias may be small in practice but finding an effective FCCA for sparsely observed functional data is still of interest and remains an open problem.

Other Functional Correlation Measures. The regularization problems for FCCA have motivated the study of alternative notions of functional correlation. These include singular correlation and singular expansions of paired processes (X, Y). While the first correlation coefficient in FCCA can be viewed as $\rho_{\text{FCCA}} = \sup_{\|u\|=\|v\|=1} \text{corr}(\langle u, X\rangle, \langle v, Y\rangle)$, observing that it is the correlation that induces the inverse problem, one could simply replace the correlation by covariance, i.e., obtain project functions u_1, v_1 that attain $\sup_{\|u\|=\|v\|=1} \mathrm{cov}(\langle u, X\rangle, \langle v, Y\rangle)$. Functions u_1, v_1 turn out to be the first pair of the singular basis of the covariance operator of (X, Y) (Yang, Müller & Stadtmüller 2011). This motivates to define a functional correlation as the first singular correlation

$$\rho_{\mathrm{SCA}} = \frac{\mathrm{cov}(\langle u_1, X\rangle, \langle v_1, Y\rangle)}{\sqrt{\mathrm{var}(\langle u_1, X\rangle)\,\mathrm{var}(\langle v_1, Y\rangle)}}. \qquad (7)$$

Another natural approach that also avoids the inverse problem is to define functional correlation as the cosine of the angle between functions in L^2. For this notion to be a meaningful measure of alignment of shapes, one first needs to subtract the integrals of the functions, i.e., their projections on the constant function 1, which corresponds to a "static part". Again considering pairs of processes $(X, Y) = (X_1, X_2)$ and denoting the projections on the constant function 1 by $M_k = \langle X_k, 1\rangle$, $k = 1, 2$, the remainder $X_k - M_k$, $k = 1, 2$, is the "dynamic part" for each random function. The cosine of the L^2-angle between the dynamic parts then provides a correlation measure of functional shapes. These ideas can be formalized as follows (Dubin & Müller 2005). Defining standardized curves either by $X_k^*(t) = (X_k(t) - M_k)/(\int (X_k(t) - M_k)^2 dt)^{1/2}$ or alternatively by also removing $\mu_k = EX_k$, $X_k^*(t) = (X_k(t) - M_k - \mu_k(t))/(\int (X_k(t) - M_k - \mu_k(t))^2 dt)^{1/2}$, the cosine of the angle between the standardized functions is $\rho_{k,l} = E\langle X_k^*, X_l^*\rangle$. The resulting dynamic correlation and other notions of functional correlation can also be extended to obtain a precision matrix for functional data. This approach has been developed by Opgen-Rhein & Strimmer (2006) for the construction of a graphical model for gene time course data.

3.2. Functional Regression

Functional regression is an active area of research and the approach depends on whether the responses or covariates are functional or vector data and include combinations of (i) functional responses with functional covariates, (ii) vector responses with functional covariates, and (iii) functional responses with vector covariates. An approach for (i) was introduced by Ramsay & Dalzell (1991) who developed the functional linear model (FLM) (15) for this case, where the basic idea already appears in Grenander (1950), who derives this as the regression of one Gaussian process on another. This model can be viewed as an extension of the traditional multivariate linear model that associates vector responses with vector covariates. The topic that has been investigated most extensively in the literature is scenario (ii) for the case where the responses are scalars and the covariates are functions. Reviews of FLMs are Müller (2005, 2011); Morris (2015). Nonlinear functional regression models will be discussed in Section 5. In the following we give a brief review of the FLM and its variants.

Functional Regression Models with Scalar Response. The traditional linear model with scalar response $Y \in \mathcal{R}$ and vector covariate $\mathbf{X} \in \mathcal{R}^p$ can be expressed as

$$Y = \beta_0 + \langle \mathbf{X}, \beta\rangle + e, \qquad (8)$$

using the inner product in Euclidean vector space, where β_0 and β contain the regression coefficients and e is a zero mean finite variance random error (noise). Replacing the vector $\mathbf{X}$ in (8) and the coefficient vector β by a centered functional covariate $X^c = X(t) - \mu(t)$

and coefficient function $\beta = \beta(t)$, for $t \in I$, one arrives at the functional linear model

$$Y = \beta_0 + \langle X^c, \beta\rangle + e = \beta_0 + \int_I X^c(t)\beta(t)dt + e, \tag{9}$$

which has been studied extensively (Cardot, Ferraty & Sarda 1999, 2003; Hall & Horowitz 2007; Hilgert, Mas & Verzelen 2013).

An ad hoc approach is to expand the covariate X and the coefficient function β in the same functional basis, such as the B-spline basis or eigenbasis in (1). Specifically, consider an orthonormal basis φ_k, $k \geq 1$, of the function space. Then expanding both X and β in this basis leads to $X(t) = \sum_{k=1}^{\infty} A_k \varphi_k(t)$, $\beta(t) = \sum_{i=1}^{\infty} \beta_k \varphi_k(t)$ and model (9) is seen to be equivalent to the traditional linear model (8) of the form

$$Y = \beta_0 + \sum_{k=1}^{\infty} \beta_k A_k + e, \tag{10}$$

where in implementations the sum on the r.h.s. is replaced by a finite sum that is truncated at the first K terms, in analogy to (2).

To obtain consistency for the estimation of the parameter function $\beta(t)$, one selects a sequence $K = K_n$ of eigenfunctions in (10) with $K_n \to \infty$. For the theoretical analysis, the method of sieves (Grenander 1981) can be applied, where the Kth sieve space is defined to be the linear subspace spanned by the first $K = K_n$ components. In addition to the basis-expansion approach, a penalized approach using either P-splines or smoothing splines has also been studied (Cardot, Ferraty & Sarda 2003). For the special case where the basis functions φ_k are selected as the eigenfunctions ϕ_k of X, the basis representation approach in (8) is equivalent to conducting a principal component regression, albeit with an increasing number of principal components. In this case, however, the basis functions are estimated rather than pre-specified, and this adds an additional twist to the theoretical analysis.

The simple functional linear model (9) can be extended to multiple functional covariates $X_1, \ldots, X_p$, also including additional vector covariates $\mathbf{Z} = (Z_1, \ldots, Z_q)$, where $Z_1 = 1$, by

$$Y = \langle \mathbf{Z}, \theta\rangle + \sum_{j=1}^{p} \int_{I_j} X_j^c(t)\beta_j(t)dt + e, \tag{11}$$

where I_j is the interval where X_j is defined. In theory, these intervals need not be the same. Although model (11) is a straightforward extension of (9), its inference is different due to the presence of the parametric component θ. A combined least squares method to estimate θ and β_j simultaneously in a one step or profile approach (Hu, Wang & Carroll 2004), where one estimates θ by profiling out the nonparametric components β_j, is generally preferred over an alternative back-fitting method. Once the parameter θ has been estimated, any approach that is suitable and consistent for fitting the functional linear model (9) can easily be extended to estimate the nonparametric components β_k by applying it to the residuals $Y - \langle\hat{\theta}, \mathbf{Z}\rangle$.

Extending the linear setting with a single index $\int_I X^c(t)\beta(t)dt$ to summarize each functional covariate, a nonlinear link function g can be added in (9) to create a functional generalized linear model (either within the exponential family or a quasi-likelihood framework and a suitable variance function)

$$Y = g(\beta_0 + \int_I X^c(t)\beta(t)dt) + e. \tag{12}$$

This Generalized Functional Linear Model has been considered when g is known (James 2002; Cardot, Ferraty & Sarda 2003; Cardot & Sarda 2005; Wang, Qian & Carroll 2010; Dou, Pollard & Zhou 2012) and when it is unknown (Müller & Stadtmüller 2005; Chen, Hall & Müller 2011). When g is unknown and the variance function plays no role, the special case of a single-index model has further been extended to multiple indices, the number of which is possibly unknown. Such "multiple functional index models" typically forgo the additive error structure imposed in (9) - (12),

$$Y = g(\int_I X^c(t)\beta_1(t)dt, \ldots, \int_I X^c(t)\beta_p(t)dt, e), \tag{13}$$

where g is an unknown multivariate function on $\mathcal{R}^{p+1}$. This line of research follows the paradigm of sufficient dimension reduction approaches, which was first proposed for vector covariates as an off-shoot of sliced inverse regression (SIR) (Duan & Li 1991; Li 1991), and has been extended to functional data in Ferré & Yao (2003); Ferré & Yao (2005); Cook, Forzani & Yao (2010) and to longitudinal data in Jiang, Yu & Wang (2014).

Functional Regression Models with Functional Response. For a function Y on I_Y and a single functional covariate $X(t)$, $s \in I_X$, two major models have been considered,

$$Y(s) = \beta_0(s) + \beta(s)X(s) + e(s), \tag{14}$$

and

$$Y(s) = \alpha_0(s) + \int_{I_X} \alpha(s,t)X^c(t)dt + e(s), \tag{15}$$

where $\beta_0(s)$ and $\alpha_0(s)$ are non-random functions that play the role of functional intercepts, and $\beta(s)$ and $\alpha(s,t)$ are non-random coefficient functions, the functional slopes.

Model (14) implicitly assumes that $I_X = I_Y$ and is most often referred to as "varying-coefficient" model. Given s, $Y(s)$ and $X(s)$ follow the traditional linear model, but the covariate effects may change with time s. This model assumes that the value of Y at time s depends only on the current value of $X(s)$ and not the history $\{X(t) : t \leq s\}$ or future values, hence it is a "concurrent regression model". A simple and effective approach to estimate β is to first fit model (14) locally in a neighborhood of s using ordinary least squares to obtain an initial estimate $\tilde{\beta}(s)$, and then to smooth these initial estimates $\tilde{\beta}(s)$ across s to get the final estimate $\hat{\beta}$ (Fan & Zhang 1999). In addition to such a two-step procedure, one-step smoothing methods have been also studied (Hoover et al. 1998; Wu & Chiang 2000; Eggermont, Eubank & LaRiccia 2010; Huang, Wu & Zhou 2002), as well as hypothesis testing and confidence bands (Wu, Chiang & Hoover 1998; Huang, Wu & Zhou 2004), with review in Fan & Zhang (2008). More complex varying coefficient models include the nested model in Brumback & Rice (1998), the covariate adjusted model in Şentürk & Müller (2005), and the multivariate varying-coefficent model in Zhu, Fan & Kong (2014), among others.

Model (15) is generally referred to as functional linear model (FLM), and it differs in crucial aspects from the varying coefficient model (14): At any given time s, the value of $Y(s)$ depends on the entire trajectory of X. It is a direct extension of traditional linear models with multivariate response and vector covariates by changing the inner product from the Euclidean vector space to L^2. This model also is a direct extension of model (9) when the scalar Y is replaced by $Y(s)$ and the coefficient function β varies with s, leading to a bivariate coefficient surface. It was first studied by Ramsay & Dalzell (1991), who proposed a penalized least squares method to estimate the regression coefficient surface $\alpha(s,t)$. When $I_X = I_Y$, it is often reasonable to assume that only the history of X affects Y, i.e., that $\alpha(s,t) = 0$ for $s < t$. This has been referred to as the "historical functional linear model" (Malfait & Ramsay 2003), because only the history of the covariate is used to model the response process. This model deserves more attention.

When $X \in \mathcal{R}^p$ and $Y \in \mathcal{R}^q$ are random vectors, the normal equation of the least squares regression of Y on X is $\text{cov}(X,Y) = \text{cov}(X,X)\beta$, where β is a $p \times q$ matrix. Here a solution can be easily obtained if $\text{cov}(X,X)$ is of full rank so its inverse exists. An extension of the normal equation to functional X and Y is straightforward by replacing covariance matrices by their corresponding covariance operators. However, an ill-posed problem emerges for the functional normal equations. Specifically, if for paired processes (X,Y) the cross-covariance function is $r_{XY}(s,t) = \text{cov}(X(s), Y(t))$ and $r_{XX}(s,t) = \text{cov}(X(s), X(t))$ is the auto-covariance function of X, we define the linear operator, $R_{XX} : L^2 \times L^2 \to L^2 \times L^2$ by $(R_{XX}\beta)(s,t) = \int r_{XX}(s,w)\beta(w,t)dw$. Then a "functional normal equation" takes the form (He, Müller & Wang 2000)

$$r_{XY} = R_{XX}\beta, \text{ for } \beta \in L^2(I_X \times I_X).$$

Since R_{XX} is a compact operator in L^2, its inverse is not bounded, leading to an ill-posed problem. Regularization is thus needed in analogy to the situation for FCCA described in Section 3.1 (He, Müller & Wang 2003). The functional linear model (9) is similarly ill-posed, however not the varying coefficient model (14), because the normal equation for the varying-coefficient model can be solved locally at each time point and does not involve inverting an operator.

Due to the ill-posed nature of the functional linear model, the asymptotic behavior of the regression estimators varies in the three design settings. For instance, a $\sqrt{n}$ rate is attainable under the varying-coefficient model (14) for completely observed functional data or dense functional data possibly contaminated with measurement errors, but not for the other two functional linear models (9) and (15) unless the functional data can be represented by a finite number of basis functions. The convergence rate for (9) depends on how fast the eigenvalues decay to zero and on regularity assumptions for β (Cai & Hall 2006; Hall & Horowitz 2007), even when functional data are observed continuously without error. An interesting phenomenon is that prediction for model (9) follows a different paradigm in which $\sqrt{n}$ convergence is attainable if the predictor X is sufficiently smooth and the eigenvalues of predictor processes are well behaved (Cai & Hall 2006). Estimation for β and asymptotic theory for model (15) were explored in Yao, Müller & Wang (2005b); He et al. (2010) for sparse functional data.

As with scalar responses, both the varying coefficient model (14) and functional linear model (15) can accommodate vector covariates and multiple functional covariates. Since each component of the vector covariate can be treated as a functional covariate with a constant value, we only discuss the extension to multiple functional covariates, $X_1, \ldots, X_p$, noting that interaction terms can be added as needed. The only change we need to make on the models is to replace the term $\beta(s)X(s)$ in (14) by $\sum_{j=1}^p \beta_j(s)X_j(s)$ and the term $\int_{I_X} \beta(s,t)X(t)dt$ in (15) by $\sum_{j=1}^p \int_{I_{X_j}} \beta_j(s,t)X_j(t)dt$, where I_{X_j} is the domain of X_j. If there are many predictors, a variable selection problem may be encountered, and when using basis expansions it is natural to employ a group lasso or similar constrained multiple variable selection method under sparsity or other suitable assumptions.

Generalized versions can be developed by adding a pre-specified link function g in models (14) and (15). For the case of the varying coefficient model and sparse functional data this has been investigated in Şentürk & Müller (2008) for the generalized varying coefficient model and for model (15) and dense functional data in James & Silverman (2005) for a finite number of expansion coefficients for each function. Jiang & Wang (2011) considered a setting where the link function may vary with time but the β in the index does not vary with time. The proposed dimension reduction approach in this paper expands the MAVE method by Xia et al. (2002) to functional data.

Random Effects Models. In addition to targeting fixed effects regression, the nonparametric modeling of random effects is also of interest. Here the random effects are contained in the stochastic part $e(t)$ of (14) and (15). One approach is to extend the FPCA approach of Section 2 to incorporate covariates (Cardot & Sarda 2006; Jiang, Aston & Wang 2009; Jiang & Wang 2010). These approaches are aiming to incorporate low dimensional projections of covariates to alleviate the curse of dimensionality for nonparametric procedures. One scenario where it is easy to implement covariate adjusted FPCA is the case where one has functional responses and vector covariates. One could conduct a pooled FPCA combining all data as a first step and then to use the FPCA scores obtained from the first stage to model covariate effects through a single-index model at each FPCA component (Chiou, Müller & Wang 2003). At this time, such approaches require dense functional data, as for sparse data individual FPC scores cannot be estimated consistently.

4. Clustering and classification of functional data

Clustering and classification are useful tools for traditional multivariate data analysis and are equally important yet more challenging in functional data analysis. We take daily vehicle speed trajectories at a fixed location as realizations of random functions as a motivating example for illustrating clusters of vehicle speed patterns. The data were recorded by a dual loop detector station located near Shea-Shan tunnel on National Highway 5 in Taiwan for 76 days during July–September 2009. The vehicle speed measures (km/hour) were averaged over 5-minute intervals. Figure 5 displays the patterns of vehicle speed for two clusters obtained by the k-centers subspace projection method to be described below. As indicated by Figure 6, Cluster 1 characterizes holidays while Cluster 2 pinpoints weekdays, reflecting the traffic patterns at the location.

In the terminology of machine learning, functional data clustering is an unsupervised learning process while functional data classification is a supervised learning procedure. Cluster analysis aims to group a set of data such that data objects within clusters are more similar than across clusters with respect to a metric. In contrast, classification assigns a new data object to a pre-determined group by a discriminant function or classifier. Functional classification typically involves training data containing a functional predictor with an associated multi-class label for each data object. The discrimination procedure of functional classification is closely related to functional cluster analysis, even though the goals are different. When the structures or centers of clusters can be established in functional data clustering, the criteria used for identifying clusters can also be used for classification. Methodology for clustering and classification of functional data has advanced rapidly during the past decades, due to rising demand for such methods in data applications. Given the vast literature on functional clustering and classification, we focus in the following on only

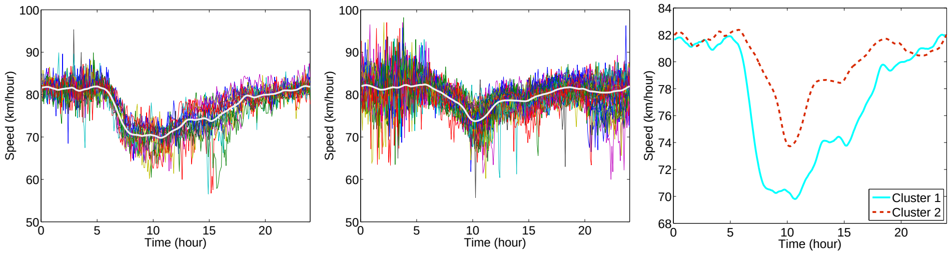

Figure 5

Observations superimposed on the estimated mean functions of daily vehicle speed recorded by a dual loop vehicle detector station for holidays (left panel for Cluster 1) and non-holidays (middle panel for Cluster 2), and the estimated cluster-specific mean functions (right panel) for comparison.

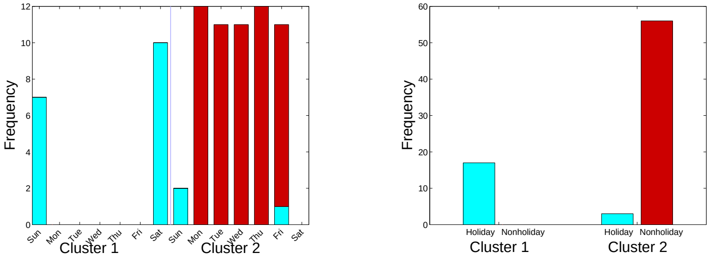

Figure 6

Histograms of cluster labels grouped by days of the week (left panel) and by (non-)holidays (right panel), respectively. Of the 76 days, 20 belong to Cluster 1 and 56 to Cluster 2.

a few typical methods.

4.1. Clustering of functional data

For vector-valued multivariate data, hierarchical clustering and the k-means partitioning methods are two classical and popular approaches. Hierarchical clustering is an algorithmic approach, using either agglomerative or divisive strategies, that requires a dissimilarity measure between sets of observations, which informs which clusters should be combined or when a cluster should be split. In the k-means clustering method, the basic idea hinges on cluster centers, the means for the clusters. The cluster centers are established through algorithms aiming to partition the observations into k clusters such that the within-cluster sum of squared distances, centering around the means, is minimized. Classical clustering concepts for vector-valued multivariate data can typically be extended to functional data, where various additional considerations arise, such as discrete approximations of distance measures, and dimension reduction of the infinite-dimensional functional data objects. In particular, k-means type clustering algorithms have been widely applied to functional data, and are more popular than hierarchical clustering algorithms. It is natural to view cluster mean functions as the cluster centers in functional clustering.

Specifically, for a sample of functional data $\{X_i(t); i = 1, \ldots, n\}$, the k-means functional clustering aims to find a set of cluster centers $\{\mu^c; c = 1, \ldots, L\}$, assuming there are L clusters, by minimizing the sum of the squared distances between $\{X_i\}$ and the cluster centers that are associated with their cluster labels $\{C_i;\ i = 1, \ldots, n\}$, for a suitable functional distance d. That is, the n observations $\{X_i\}$ are partitioned into L groups such that

$$\frac{1}{n}\sum_{i=1}^{n} d^2(X_i, \mu_n^c), \tag{16}$$

is minimized over all possible sets of functions $\{\mu_n^c;\ c = 1, \ldots, L\}$, where $\mu_n^c(t) = \sum_{i=1}^{n} X_i(t)\mathbf{1}_{\{C_i=c\}}/N_c$, and $N_c = \sum_{i=1}^{n} \mathbf{1}_{\{C_i=c\}}$. The distance d is often chosen as the L^2 norm. Since functional data are discretely recorded, frequently contaminated with measurement errors, and can be sparsely or irregularly sampled, a common approach to minimize (16) is to project the infinite-dimensional functional data onto a low dimensional space of a set of basis functions, similarly to functional correlation and regression.

There is a vast amount of literature on functional data clustering during the past decade, including methodological development and a broad range of applications. Some selected approaches to be discussed below include k-means type clustering in Section 4.1.1, subspace projected clustering methods in Section 4.1.2 and model-based functional clustering approaches in Section 4.1.3.

4.1.1. Mean functions as cluster centers.

The traditional k-means clustering for vector-valued multivariate data has been extended to functional data using mean functions as cluster centers, and one can distinguish two typical approaches, as follows.

Functional Clustering via Functional Basis Expansion. As described in Section 2, given a set of pre-specified basis functions $\{\varphi_1, \varphi_2, \ldots\}$ of the function space, the first K projections $\{B_{ik}\}$ of the observed trajectories onto the space spanned by the set of basis functions can be used to represent the functional data, where $B_{ik} = \langle X_i^c, \varphi_k\rangle$, $k = 1, \ldots, K$. The distributional patterns of $\{B_{ik}\}$ then reflect the clusters in function space. Therefore, a typical functional clustering approach via functional basis expansion is to represent the functional data by the set of coefficients in the basis expansion, which requires a judicious choice of the basis functions, and then applying available clustering algorithms for multivariate data, such as the k-means algorithm, to partition the estimated sets of coefficients. When clustering the fitted sets of coefficients $\{B_{ik}\}$ with the k-means algorithm, one obtains cluster centers $\{\bar{B}_1^c, \ldots, \bar{B}_K^c\}$ on the projected space, and thus the set of cluster centers in the function space $\{\hat{\mu}^c;\ c = 1, \ldots, L\}$, where $\hat{\mu}^c(t) = \sum_{k=1}^{K} \bar{B}_k^c\varphi_k(t)$.

Such two stage clustering has been adopted in Abraham et al. (2003) using B-spline basis functions and Serban & Wasserman (2005) using Fourier basis functions coupled with the k-means algorithm, as well as Garcia-Escudero & Gordaliza (2005) using B-splines with a robust trimmed k-means method. Abraham et al. (2003) derived the strong consistency property of this clustering method, which has been implemented with various basis functions, such as P splines (Coffey, Hinde & Holian 2014), a Gaussian orthonormalized basis (Kayano, Dozono & Konishi 2010), and the wavelet basis (Giacofci et al. 2013).

Functional Clustering via FPCA. In contrast to the functional basis expansion approach that requires a pre-specified set of basis functions, the finite approximation FPCA approach using the FPCs as described in Section 2 employs data-adaptive basis functions that are determined by the covariance function of the functional data. Then the distributions of the sets of FPC scores $\{A_{ik}\}$ (see (2)) indicate different cluster patterns, while the overall mean function $\mu(t)$ does not affect clustering, and the scores $\{A_{ik}\}$ play a similar role as the basis coefficients $\{B_{ik}\}$ for clustering. Peng & Müller (2008) used a k-means algorithm on the FPC scores, employing a special distance adapted to clustering sparse functional data, and Chiou & Li (2007) used a k-means algorithm on the FPC scores as an initial clustering step for the subspace projected k-centers functional clustering algorithm. When the mean functions as the cluster centers are sufficient to define the clusters, this step is sufficient. However, when covariance structures also play a role to distinguish clusters, taking mean functions as cluster centers is not adequate, as will be discussed in the next subsection.

4.1.2. Subspaces as cluster centers. Since functional data are realizations of random functions, it is natural to use differences in the stochastic structure of random functions for clustering. This idea is particularly sensible in functional data clustering, utilizing the Karhunen-Loève representation in (1). More specifically, the truncated representation (2) of random functions in addition to the mean includes the sum of a series of linear combinations of the eigenfunctions of the covariance operator with the FPC scores as the weights. The subspace spanned by the components of the expansion, the mean function and the set of the eigenfunctions, can be used to characterize clusters. Therefore, clusters of the data set are identified via subspace projection such that cluster centers hinge on the stochastic structure of the random functions, rather than the mean functions only.

The FPC subspace-projected *k-centers functional clustering* approach was considered in Chiou & Li (2007), using subspaces as cluster centers. Let C be the cluster membership variable, and the FPC subspace $\mathcal{S}^c = \{\mu^c, \phi_1^c, \ldots, \phi_{K_c}^c\}$, $c = 1, \ldots, L$, assuming that there are L clusters. The projected function of X_i onto the FPC subspace $\mathcal{S}^c$ can be written as

$$\tilde{X}_i^c(t) = \mu^c(t) + \sum_{k=1}^{K_c} A_{ik}^c \phi_k^c(t). \tag{17}$$

One aims to find the set of cluster centers $\{\mathcal{S}^c;\, c = 1, \ldots, L\}$, such that the best cluster membership of X_i, $c^*(X_i)$, is determined by minimizing the discrepancy between the projected function $\tilde{X}_i^c$ and the observation X_i,

$$c^*(X_i) = \underset{c \in \{1,\ldots,L\}}{\arg\min} \sum_{i=1}^{n} d^2(X_i, \tilde{X}_i^c). \tag{18}$$

In contrast, k-means clustering aims to find the set of cluster sample means as the cluster centers, instead of the subspaces spanned by $\{\mathcal{S}^c; c = 1, \ldots, L\}$. The initial step of the subspace-projected clustering procedure uses only μ^c, which reduces to the k-means functional clustering. In subsequent iteration steps, the mean function and the set of eigenfunctions for each cluster is updated and used to identify the set of cluster subspaces $\{\mathcal{S}^c\}$, until iterations converge. This functional clustering approach simultaneously identifies the structural components of the stochastic representation for each cluster. The idea of the k-centers function clustering via subspace projection was further developed to clustering functional data with similar shapes based on a shape function model with random scaling effects (Chiou & Li 2008).

More generally, in probabilistic clustering the cluster membership of X_i may be determined by maximizing the conditional cluster membership probability given X_i, $P_{C|X}(c \mid X_i)$, such that

$$c^*(X_i) = \underset{c \in \{1,\ldots,L\}}{\arg\max} P_{C|X}(c \mid X_i). \tag{19}$$

This criterion requires modeling of the conditional probability $P_{C|X}(\cdot \mid \cdot)$. It can be achieved by a generative approach that requires a joint probability model or alternatively through a discriminative approach using, for example, a multi-class logit model (Chiou 2012).

For the k-means type or the k-centers functional clustering algorithms, the number of clusters is pre-specified. The number of clusters for subspace projected functional clustering can be determined by finding the maximum number of clusters while retaining significant differences between pairs of cluster subspaces. Li & Chiou (2011) developed the forward functional testing procedure to identify the total number of clusters under the framework of subspace projected functional data clustering.

4.1.3. Functional clustering with mixture models. Model-based clustering (Banfield & Raftery 1993) based on mixture models is widely used in clustering vector-valued multivariate data and has been extended to functional data clustering. In this approach, the mixture model determines the cluster centers. Similarly to the k-means type of functional data clustering, typical mixture model-based approaches to functional data clustering in a first step project the infinite dimensional functional data onto low-dimensional subspaces. An example is James & Sugar (2003), who applied functional clustering models based on Gaussian mixture distributions to the natural cubic spline basis coefficients, with emphasis on clustering sparsely sampled functional data. Similarly, Jacques & Preda (2013, 2014) applied the idea of Gaussian mixture modeling to FPCA scores. All these methods are based on truncated expansions as in (2).

Random effects modeling also provides a model-based clustering approach, using mixed effects models with B-splines or P-splines, for example to cluster time-course gene expression data (Coffey, Hinde & Holian 2014). For clustering longitudinal data, a linear mixed model for clustering using a penalized normal mixture as random effects distribution has been studied (Heinzl & Tutz 2014). Bayesian hierarchical clustering also plays an important role in the development of model-based functional clustering, typically assuming Gaussian mixture distributions on the sets of basis coefficients fitted to individual trajectories. Dirichlet processes are frequently used as prior for the mixture distributions and also to deal with the uncertainty in the cluster numbers (Angelini, De Canditiis & Pensky 2012; Rodriguez, Dunson & Gelfand 2009; Petrone, Guindani & Gelfand 2009; Heinzl & Tutz 2013).

4.2. Classification of functional data

While functional clustering aims at finding clusters by minimizing an objective function such as (16) and (18), or more generally, by maximizing the conditional probability as in (19), functional classification assigns a group membership to a new data object with a discriminant function or a classifier. Popular approaches for functional data classification are based on functional regression models that feature class labels as responses and the observed functional data and other covariates as predictors. This leads to regression based functional data classification methods, for example, functional generalized linear regression models and functional multiclass logit models. Similar to functional data clustering, most functional data classification methods apply a dimension reduction technique using a truncated expansion in a pre-specified function basis or in the data-adaptive eigenbasis.

4.2.1. Functional regression for classification. For regression-based functional classification models, functional generalized linear models (James 2002; Müller 2005; Müller &

Stadtmüller 2005) or more specifically, functional binary regression, such as functional logistic regression, are popular approaches. For a random sample $\{(Z_i, X_i); i = 1, \dots, n\}$, where Z_i represents a class label, $Z_i \in \{1, \dots, L\}$ for L classes, associated with functional observations X_i, a classification model for an observation X_0 based on functional logistic regression is

$$\log \frac{Pr(Z = k \mid X_0)}{Pr(Z_i = L \mid X_0)} = \gamma_{0k} + \int_{\mathcal{T}} X_0(t)\gamma_{1k}(t)dt, \quad k = 1, \dots, L-1, \tag{20}$$

where γ_{0k} is an intercept term and $\gamma_{1k}(t)$ the coefficient function of the predictor $X_0(t)$ and $Pr(Z_i = L \mid X_i) = 1 - \sum_{k=1}^{L} Pr(Z_i = k \mid X_i)$. This is a functional extension of the baseline odds model in multinomial regression (McCullagh & Nelder 1983).

Given a new observation X_0, the model-based Bayes classification rule is to choose the class label Z_0 with the maximal posterior probability among $\{Pr(Z_0 = k \mid X_0); k = 1, \dots, L\}$. More generally, Leng & Müller (2006) used the generalized functional linear regression model based on the FPCA approach. When the logit link is used in the model, it becomes the functional logistic regression model, several variants of which have been studied (Araki et al. 2009; Matsui, Araki & Konishi 2011; Wang, Ray & Mallick 2007; Zhu, Vannucci & Cox 2010; Rincon & Ruiz-Medina 2012).

4.2.2. Functional discriminant analysis for classification. In contrast to the regression-based functional classification approach, another popular approach is based on the classical linear discriminant analysis method. The basic idea is to classify according to the largest conditional probability of the class label variable given a new data object by applying the Bayes rule. Suppose that the kth class has prior probability π_k, $\sum_{k=1}^{K} \pi_k = 1$. Given the density of the kth class, f_k, the posterior probability of a new data object X_0 is given by the Bayes formula,

$$Pr(Z = k \mid X_0) = \frac{\pi_k f_k(X_0)}{\sum_{j=1}^{K} \pi_j f_j(X_0)}. \tag{21}$$

Developments along these lines include a functional linear discriminant analysis approach to classify curves (James & Hastie 2001), a functional data-analytic approach to signal discrimination, using the FPCA method for dimension reduction (Hall, Poskitt & Presnell 2001) and kernel functional classification rules for nonparametric curve discrimination (Ferraty & Vieu 2003; Chang, Chen & Ogden 2014; Zhu, Brown & Morris 2012). Theoretical support and a notion of "perfect classification" standing for asymptotically vanishing misclassification probabilities has been introduced in Delaigle & Hall (2012) for linear and Delaigle & Hall (2013) for quadratic functional classification.

5. Nonlinear Methods for Functional Data

Due to the complexity of functional data analysis, which blends stochastic process theory, functional analysis, smoothing and multivariate techniques, most research at this point has focused on linear functional models, such as functional principal components and functional linear regression, which are reviewed in Sections 2 and 3. Perhaps owing to the success of these linear approaches, the development of nonlinear methods has been slower, even though in many situations linear methods are not fully adequate. A case in point is the presence of time variation or time warping that has been observed for many functional data (Kneip & Gasser 1992; Gasser & Kneip 1995). This means that observation time itself is randomly distorted and time variation may constitute the main source of variation for some functional data (Wang & Gasser 1997; Ramsay & Li 1998). Efficient models will then need to reflect the nonlinear features of the data.

5.1. Nonlinear Regression Models

The classical functional regression models are linear models, as described in Section 3.2, see equations (9), (9), (11), (15). Direct nonlinear extensions still contain a linear predictor, but combine it with a nonlinear link function, in a similar fashion as the generalized linear model (McCullagh & Nelder 1983). These are the generalized functional linear model and single index models (12) and (13). From a theoretical perspective, the presence of a nonlinear link functions complicates the analysis of these models, e.g., requiring to decompose such models into a series of $p-$dimensional approximation models with $p \to \infty$ (Müller & Stadtmüller 2005).

There have been various developments towards fully nonparametric regression models for functional data (Ferraty & Vieu 2006), which lie at the other end of the spectrum in comparison to the functional linear model. These models extend the concept of nonparametric smoothing to the case of predictor functions, where for scalar responses Y one considers functional predictors X, aiming at $E(Y \mid X) = g(X)$ for a smooth regression function g, for example extending kernel smoothing to this situation. The idea is to replace differences in the usual Euclidean predictor space by a projected pseudo-distance in a functional predictor space, so that the scaled kernel $K(\frac{x-y}{h})$ with a bandwidth h becomes $K(\frac{d(x,y)}{h})$, where d is a metric in the predictor space (Ferraty & Vieu 2006). Due to the infinite nature of the predictors, when choosing d as the L^2 distance, such models are subject to a serious form of "curse of dimensionality", as functional predictors are inherently infinite-dimensional. This "curse" is quantifiable in terms of unfavorable small ball probabilities in function space (Delaigle & Hall 2010). What this means is that an appropriate choice of the metric d that avoids the "curse" is essential, and whether such a choice is possible for a given functional data set and how to implement remains an open problem. In some cases, when data are clustered in lower-dimensional manifolds, the rates of convergence pertaining to the lower dimension will apply (Bickel & Li 2007), counteracting the "curse".

More generally, to bypass the "curse" and the metric selection problem, it is of interest to consider nonlinear models, which are subject to some structural constraints that do not overly infringe flexibility. One can aim at models that retain polynomial rates of convergence, while they are more flexible than, say, functional linear models. Such models are particularly useful when diagnostics for the functional linear model indicate lack of fit (Chiou & Müller 2007). An example are generalized functional linear models (12) as well as extensions to single index models (Chen, Hall & Müller 2011) that provide enhanced flexibility and structural stability while model fits converge at polynomial rates.

Additive Models for Functional Data. Beyond single index models, another powerful dimension reduction tool is the additive model (Stone 1985; Hastie & Tibshirani 1986), which also has been extended to functional data (Lin & Zhang 1999; You & Zhou 2007; Carroll et al. 2009; Lai, Huang & Lee 2012). In these models it is generally assumed that the time effect is additive, which is sometimes restrictive. Modeling additive components that are bivariate functions of time and a covariate (Zhang, Park & Wang 2013), this restriction can be avoided. The two-dimensional smoothing needed for the bivariate functions each component may be replaced by one-dimensional smoothing steps, if one further assumes that each of the additive components is the product of an unknown time effect and an unknown covariate effect (Zhang & Wang 2015), which leads to easy interpretation and implementation.

Alternatively, one can utilize the functional principal components A_k, as defined in (1), for dimension reduction of the predictor process or processes X, and then assume that the regression relation is additive in these components. While the linear functional regression model with scalar response can be written as $E(Y \mid X) = EY + \sum_{k=1}^{\infty} A_k \beta_k$ with an infinite sequence of regression coefficients β_k, the *functional additive model* is

$$E(Y \mid X) = EY + \sum_{k=1}^{\infty} f_k(A_k), \tag{22}$$

where the component functions are required to be smooth and to satisfy $E(f_k(A_k)) = 0$ (Müller & Yao 2008; Sood, James & Tellis 2009).

This model can be characterized as frequency-additive. A key feature that makes this model not only easy to implement but also accessible to asymptotic analysis, is

$$E(Y - \mu_Y \mid A_k) = E\{E(Y - \mu_Y|X) \mid A_k\} = E\{\sum_{j=1}^{\infty} f_j(A_j) \mid A_k\} = f_k(A_k), \tag{23}$$

if the functional principal components A_k are assumed to be independent, as would be the case for Gaussian predictor processes, where $\mu_Y = EY$. This implies that simple one-dimensional smoothing of the responses against the FPCs leads to consistent estimates of the component functions f_k (Müller & Yao 2008), so that the usual backfitting that is normally required for additive modeling is not needed. For the functional linear model (9), already uncorrelatedness of the FPCs of the predictor processes suffices for the representation $E(Y - \mu_Y \mid A_k) = \beta_k A_k$, motivating to decompose functional linear regression into a series of simple linear regressions (Chiou & Müller 2007; Müller et al. 2009).

Projections on a finite number of directions for each of potentially many predictor functions provide an alternative additive approach that is of practical interest when the projections are formed by taking into consideration the relation between X and Y, in contrast to other functional regression models, where the predictors are formed merely based on the auto-covariance structure of predictor processes X (James & Silverman 2005; Chen, Hall & Müller 2011; Fan et al. 2014).

Still other forms of additive models have been considered for functional data. While model (22) can be characterized as frequency-additive, as it is additive in the FPCs, one may ask the question whether there are time-additive models. It is immediately clear that since the number of time points on an interval domain is uncountable, an unrestricted time-additive model $E(Y \mid X) = \sum_{t \in [0,T]} f_t(X(t))$ is not feasible. Addressing this conundrum by assuming that the functions f_t are smoothly varying in t and considering a sequence of time-additive models on increasingly dense finite grids of size p leads to the sequence

$$E(Y|X(t_1), \ldots, X(t_p)) = \sum_{j=1}^{p} f_j(X(t_j)),$$

where $f_j(x) = g(t_j, x)$ for a smooth bivariate function g. In the limit $p \to \infty$ this becomes the *continuously additive model* (Müller, Wu & Yao 2013)

$$E(Y|X) = \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} g(t_j, X(t_j)) = \int_{[0,T]} g(t, X(t))\, dt. \tag{24}$$

This model can be implemented with a bivariate spline representation of the function g; a very similar model was introduced under the name *functional generalized additive model* in McLean et al. (2014). Nonlinear or linear models, where individual predictor times are better predictors than functional principal components, i.e., regression models with time-based rather than frequency-based predictors, can be viewed as special cases of the continuously additive model (24), where only a few time points and their associated additive functions $f_j(X(t_j))$ are assumed to be predictive (Ferraty, Hall & Vieu 2010).

Optimization and Gradients With Functional Predictors. In some applications one aims to maximize the response $E(Y \mid X)$ in terms of the predictor function X. An example is the evolution of reproductive trajectories X in medflies, measured in terms of daily egg-laying, where evolution may work to maximize a desirable outcome such as lifetime reproductive success Y, as this conveys an evolutionary advantage. While enhanced egg-laying at all ages enhances lifetime reproductive success, measured as total number of eggs produced during the lifetime of a fly, it also promotes mortality, through the "cost of reproduction". However, shorter lifespan implies reduced total number of eggs. The optimal egg-laying trajectory is therefore not simply the maximal egg-laying possible at all ages but a complex trade-off between maximizing daily egg-laying and the cost of reproduction in terms of mortality (Müller et al. 2001).

To address the corresponding functional maximization problem, gradients with respect to functional predictors X are of interest (Hall, Müller & Yao 2009). Extending the functional additive model (22), one can introduce additive gradient operators with arguments in L^2 at each predictor level $X \equiv \{A_1, A_2, \ldots\}$,

$$\Gamma_X^{(1)}(u) = \sum_{k=1}^{\infty} f_k^{(1)}(A_k) \int \phi_k(t)u(t)dt, \quad u \in L^2. \tag{25}$$

These additive gradient operators serve to find directions in which responses increase, thus enabling a maximal descent algorithm in function space (Müller & Yao 2010a).

Polynomial Functional Regression. Finally, just as the common linear model can be embedded in a more general polynomial version, a polynomial functional model that extends the functional linear model has been proposed (Yao & Müller 2010), with quadratic functional regression as the most prominent special case. With centered predictor processes X^c, this model can be written as

$$E(Y \mid X) = \alpha + \int \beta(t)X^c(t)dt + \int\int \gamma(s,t)X^c(s)X^c(t)dsdt, \tag{26}$$

and in addition to the parameter function β that it shares with the functional linear model it also features a parameter surface γ. The extension to higher order polynomials is obvious. These models can be equivalently represented as polynomials in the corresponding FPCs. A natural question is whether the linear model is sufficient or needs to be extended to a model that includes a quadratic term. A corresponding test was developed by Horváth & Reeder (2013).

5.2. Time Warping, Dynamics and Manifold Learning for Functional Data

In addition to amplitude variation, many functional data are best described by assuming that additional time variation is present, i.e, the time axis is distorted by a smooth random process. A classical example are growth data. In human growth, the biological age of different children varies and this variation has a direct bearing on the growth rate that follows a general shape, but with subject-specific timing, which manifests itself for example in the subject-specific timing of the two major growth spurts, the pubertal and the pre-pubertal growth spurt (Gasser et al. 1984).

Time Variation and Curve Registration. If both amplitude and time variation are present in functional data, they cannot be separately identified, so additional assumptions that break the non-identifiability are crucial if one wishes to identify and separate these two components, which jointly generate the observed variation in the data. An example when time warping is identifiable is briefly discussed in the next section. In the presence of time warping, which is also known as curve registration or curve alignment problem, cross-sectional mean functions are inefficient and uninterpretable, because if important features such as peak locations randomly vary from curve to curve, ignoring the differences in timing when taking a cross-sectional mean will distort these features. Then the mean curve will not resemble any of the sample curves and is not useful as a representative of the sample of curves (Ramsay & Li 1998).

Early approaches to time-warped functional data included dynamic time warping (Sakoe & Chiba 1978; Wang & Gasser 1997) for the registration of speech and self-modeling nonlinear regression (Lawton & Sylvestre 1971; Kneip & Gasser 1988), where in the simplest case one assumes that the observed random functions can be modeled as shift-scale family of an unknown template function, where shift and scale are subject-specific random variables. A traditional method for time-warped functional data is the landmark method. Special features such as peak locations in functions or derivatives are aligned to their average location and then smooth transformations from the average location to the location of the feature for a specific subject are introduced (Kneip & Gasser 1992; Gasser & Kneip 1995). If well-expressed features are present in all sample curves, the landmark method serves as a gold standard for curve alignment. A problem is that landmarks may be missing in some sample functions or may be hard to identify due to noise in the data.

The mapping of latent bivariate time warping and amplitude processes into random functions has been studied systematically, leading to the definition of the mean curve as the function that corresponds to the bivariate Fréchet mean of both time warping and amplitude processes (Liu & Müller 2004), which can be exemplified with relative area-under-the curve warping, where the latter has been shown to be particularly well suited for samples of random density functions (Kneip & Utikal 2001; Zhang & Müller 2011)

Recent approaches include registration by minimizing a Fisher-Rao metric (Tucker, Wu & Srivastava 2013; Wu & Srivastava 2014), alignment of event data by dynamic time warping (Arribas-Gil & Müller 2014), and joint models for amplitude and time variation or for combinations of regression and time variation (Kneip & Ramsay 2008), where adopting a joint perspective leads to better interpretability of time warping models for spoken language (Hadjipantelis et al. 2015) or better performance of functional regression in the presence of warping (Gervini 2015).

Pairwise Warping. As a specific example of a warping approach, we discuss a pairwise warping approach that is based on the idea that all relevant information about time warping resides in pairwise comparisons and the resulting pairwise relative time warps (Tang & Müller 2008). Starting with a sample of n i.i.d. smooth observed curves $Y_1, Y_2, ..., Y_n$ (with suitable modifications for situations where the curves are not directly observed but only noisy measurements of the curves at a grid of discrete time points are available) we postulate that

$$Y_i(t) = X_i\{h_i^{-1}(t_j)\},\ t \in [0, T], \tag{27}$$

where the X_i are i.i.d. random functions that represent amplitude variation and the h_i are the realizations of a time warping process h that yields warping functions that represent time variation, are strictly monotone and invertible and satisfy $h_i(0) = 0$, $h_i(T) = T$. The time warping functions map time onto warped time. Traditionally, time is assumed to flow forward and therefore warping functions are strictly monotone increasing. However, a recent time warping approach that allows time to flow backwards has been shown to be useful for modeling declines in house prices as time reversals (Peng, Paul & Müller 2014).

To break the non-identifiability, which is a characteristic of time warping models as already mentioned, Tang & Müller (2008) make the assumptions that the overall curve variation is (at least asymptotically) dominated by time variation, i.e., $X_i(t) = \mu(t) + \delta Z_i(t)$, where δ vanishes for increasing sample size n, the Z_i are realizations of a smooth square integrable process and $E\{h(t)\} = t$, for $t \in [0, 1]$. Then warping functions may be represented in a suitable basis that ensures monotonicity and has associated random coefficients in the expansion, for example monotonically restricted piecewise linear functions. If curve Y_i has the associated time warping function h_i then the warping function g_{ik} that transforms the time scale of curve Y_i towards that of Y_k is $g_{ik}(t) = h_i\{h_k^{-1}(t)\}$, and analogously, the pairwise-warping function of curve Y_k towards Y_i is $g_{ki}(t) = h_k\{h_i^{-1}(t)\}$.

Because warping functions are assumed to have average identify, $E[h_i\{h_k^{-1}(t)\}\big|h_k] = h_k^{-1}(t)$, and, as $g_{ik}(t) = h_i\{h_k^{-1}(t)\}$, we find that $h_k^{-1}(t) = E\{g_{ik}(t)\big|h_k\}$, which motivates corresponding estimators by plugging in estimates of the pairwise warping functions. This shows that under certain regularity assumptions the relevant warping information is indeed contained in the pairwise time warpings.

Functional Manifold Learning. A comprehensive approach to time warping and other nonlinear features of functional data such as scale or scale-shift families that simultaneously handles amplitude and time warping features is available through manifold learning. A motivation for the use of functional manifold models is that image data that are dominated by random domain shifts lie on a manifold (Donoho & Grimes 2005). Similar warping models where the warping component corresponds to a random time shift have been studied for functional data (Silverman 1995; Leng & Müller 2006). Such data have low-dimensional representations in a transformed space but are infinite-dimensional in the traditional functional basis expansion including the eigenbasis expansion (1). While these expansions will always converge in L^2 under minimal conditions, they are inefficient in comparison with representations that take advantage of the manifold structure.

When functional data include time warping or otherwise lie on a nonlinear low-dimensional manifold that is situated within the ambient infinite-dimensional functional Hilbert space, desirable low-dimensional representations can be obtained through manifold learning; the resulting nonlinear representations are particularly useful for subsequent statistical analysis. Once a map from an underlying low-dimensional Euclidean space into functional space has been determined, this gives the desired manifold representation. Among the various nonlinear dimension reduction methods that employ manifold learning (Roweis & Saul 2000), Isomap (Tenenbaum, De Silva & Langford 2000) can be easily implemented and has been shown to be a useful and versatile method for functional data analysis. Specifically, a modified Isomap learning algorithm that adds a penalty to the empirical geodesic distances to correct for noisy data, and employs kernel smoothing to map data from the manifold into functional space provides a flexible and broadly applicable approach to low-dimensional manifold modeling of time-warped functional data (Chen & Müller 2012). This approach targets "simple" functional manifolds $\mathcal{M}$ in L^2 that are "flat", i.e., isomorphic to a subspace of Euclidean space, such as a Hilbert space version of the "Swiss Roll". An essential input for Isomap is the distance between functional data. The default distance is the L^2 distance in function space, but this distance is not always feasible, for example when functional data are only sparsely sampled. In such cases, the L^2 distance needs to be replaced by a distance that adjusts to sparsity (Peng & Müller 2008).

The manifold $\mathcal{M}$ is characterized by a coordinate map $\varphi : \mathbb{R}^d \to \mathcal{M} \subset L^2$, such that φ is bijective, and both φ, φ^{-1} are continuous and isometric. For a random function X the mean μ in the d-dimensional representation space and the manifold mean $\mu^{\mathcal{M}}$ in the functional L^2 space are characterized by

$$\mu = \mathrm{E}\{\varphi^{-1}(X)\}, \quad \mu^{\mathcal{M}} = \varphi^{-1}(\mu).$$

The isometry of the map φ implies that the manifold mean $\mu^{\mathcal{M}}$ is uniquely defined.

In addition to obtaining a mean, a second basic task in FDA is to quantify variation. In analogy to the modes of variation that are available through eigenfunctions and FPCA (Castro, Lawton & Sylvestre 1986; Jones & Rice 1992), one can define *manifold modes of variation*

$$X_{j,\alpha}^{\mathcal{M}} = \varphi\big(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}}\mathbf{e}_j^{\mathcal{M}}\big),\ j = 1, \ldots, d,\ \alpha \in \mathbb{R},$$

where the vectors $\mathbf{e}_j^{\mathcal{M}} \in \mathbb{R}^d$, $j = 1, \ldots, d$, are the eigenvectors of the covariance matrix of $\varphi^{-1}(X) \in \mathbb{R}^d$, i.e., $\mathrm{cov}(\varphi^{-1}(X)) = \sum_{j=1}^d \lambda_j^{\mathcal{M}}(\mathbf{e}_j^{\mathcal{M}})(\mathbf{e}_j^{\mathcal{M}})^T$. Here $\lambda_1^{\mathcal{M}} \geq \ldots \geq \lambda_d^{\mathcal{M}}$ are the corresponding eigenvalues and the modes are represented by varying the scaling factors α.

Each random function $X \in \mathcal{M}$ then has a unique representation in terms of the $d-$dimensional vector $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_d) \in \mathbb{R}^d$,

$$X = \varphi(\boldsymbol{\mu} + \sum_{j=1}^d \vartheta_j \mathbf{e}_j^{\mathcal{M}}), \quad \vartheta_j = \langle \varphi^{-1}(X) - \boldsymbol{\mu}, \mathbf{e}_j^{\mathcal{M}} \rangle,\ j = 1, \ldots, d,$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^d$ and ϑ_j are uncorrelated r.v.s with mean 0 and variance $\lambda_j^{\mathcal{M}}$, the functional manifold components (Chen & Müller 2012). This representation is a genuine dimension reduction of the functional data to the finite dimension d, while for functional data that lie on a nonlinear manifold the Karhunen-Loève representation usually will require a large number of components to provide a good approximation.

In practical applications, the presence of functional manifolds is often evident when plotting the second functional principal components (FPCs) versus the first FPCs, where curved shapes such as "horseshoes" will appear and typically there are few data in the vicinity of the mean (which is always 0 for all principal components, which are uncorrelated). An example is in the right upper panel of Fig. 7, where the functional principal components falling near the mean have low density, while they cluster around the periphery.

This example is based on a small sample ($n = 20$) of house price index curves showing the boom-bust cycle 2000-2009. These data were previously discussed and modeled in Peng, Paul & Müller (2014) and are displayed in the left upper panel of Fig. 7. The curves show mostly amplitude variation with some variation in the timing of the peaks, many of which have a bimodal appearance, indicating that when the boom cycle peaked, there was a first

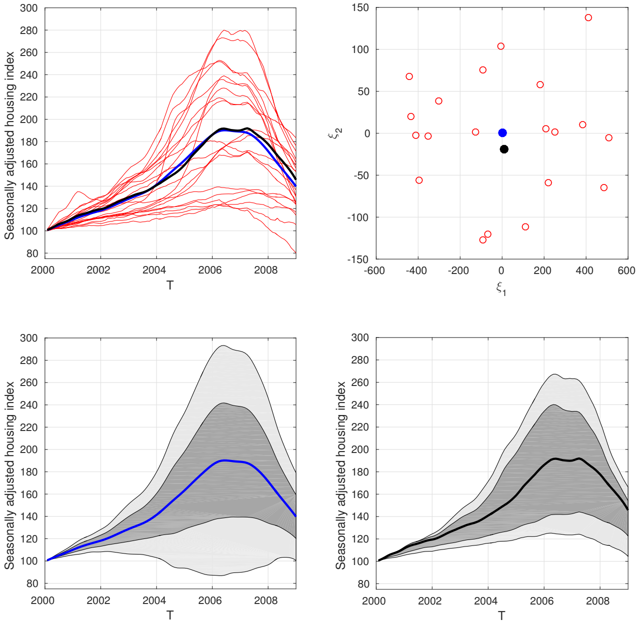

Figure 7

Left upper panel: Median housing price index trajectories for 20 U.S. metropolitan regions, 2000-2009, values in the year 2000 are standardized at 100. Blue curve is the standard cross-sectional mean, while black curve is the manifold mean. Right upper panel: Second functional principal component plotted against first functional principal component, with the regular mean in blue and the manifold mean in black. Lower left panel: Standard first mode of variation based on L^2 eigenanalysis. Lower right panel: Corresponding first manifold mode of variation.

slight downturn, followed a short period of further price increase, just before the onset of the major downturn. The manifold mean and first mode of variation in the right panel reflects this quite well, and also shows small shifts in peak locations as well as shape changes, while the ordinary mode of variation does not allow for a peak shift and shows hardly any bimodality (left lower panel). This is an example where the differences between the manifold representation and the Karhune-Loève representation are not major, presumably due to the low sample size, while the manifold representation is clearly preferrable.

Learning Dynamics From Functional Data. Since functional data consist of repeated observations of (usually) time-dynamic processes, they can be harnessed to determine the dynamics of the underlying processes. Dynamics are typically assessed with derivatives. Under regularity conditions, derivatives $X^{(\nu)}$ of square integrable processes X are also square integrable and the eigenrepresentation (1) implies

$$X_i^{(\nu)}(t) = \mu^{(\nu)}(t) + \sum_{k=1}^{\infty} A_{ik}\phi_k^{(\nu)}(t), \qquad (28)$$

where ν is the order of derivative. Derivatives of μ can be estimated with suitable smoothing methods and those of ϕ by partial differentiation of covariance surfaces, which is even possible in the case of sparsely sampled data, where direct differentiation of trajectories would not be possible (Liu & Müller 2009).

For the case where one has differentiable Gaussian processes, since X and $X^{(1)}$ are jointly Gaussian, it is easy to see that (Müller & Yao 2010b)

$$X^{(1)}(t) - \mu^{(1)}(t) = \beta(t)\{X(t) - \mu(t)\} + Z(t),\ \beta(t) = \frac{\mathrm{cov}\{X^{(1)}(t), X(t)\}}{\mathrm{var}\{X(t)\}}. \tag{29}$$

This is a linear differential equation with a time-varying function $\beta(t)$ and a drift process Z. Here Z is a Gaussian process such that $Z(t)$, $X(t)$ are independent at each t. If Z is relatively small, the equation is dominated by the linear part and the function β. Then the behavior of β characterizes different dynamics, where one can distinguish *dynamic regression to the mean* for those t where $\beta(t) < 0$ and *explosive behavior* for those t where $\beta(t) > 0$. In the first case, deviations of $X(t)$ from the mean function $\mu(t)$ will diminish, while in the second case they will increase: An individual with a value $X(t)$ above the mean will tend to move even higher above the mean under the explosive regimen but will move closer to the mean under dynamic regression to the mean. Thus the function β that is estimated from the observed functional data can be used to characterize the underlying empirical dynamics.

A nonlinear version of dynamics learning can be developed for the case of non-Gaussian processes (Verzelen, Tao & Müller 2012). This is of interest whenever linear dynamics is not applicable, and is based on the fact that one always has a function f such that

$$E\{X^{(1)}(t) \mid X(t)\} = f\{t, X(t)\}, \quad X^{(1)}(t) = f\{t, X(t)\} + Z(t)\ , \tag{30}$$

with $E\{Z(t) \mid X(t)\} = 0$ almost surely. Generally the function f will be unknown. It can be consistently estimated from the observed functional data by nonparametrically regressing derivatives $X^{(1)}$ against levels X and time t. This can be implemented with simple smoothing methods. The dynamics of the processes is then jointly determined by the function f and the drift process Z. Nonlinear dynamics learning is of interest to understand the characteristics of the underlying stochastic system and can also be used to determine whether individual trajectories are "on track", for example in applications to growth curves.

6. Outlook and Future Perspectives

FDA has widened its scope from a relatively narrow focus on the analysis of samples of fully observed functions to much wider applicability. An example is longitudinal data analysis, where FDA provides a rich nonparametric methodology for a field that has been dominated by parametric random effects models for a long time. Also of special interest are recent developments in the interface of high-dimensional and functional data. These include: Combining functional elements with high-dimensional covariates, such as modeling predictor times that exercise an individual predictor effect on an outcome that goes beyond the functional linear model (Kneip & Sarda 2011), predictor selection among high-dimensional functional principal component scores and baseline covariates in functional regression models (Kong et al. 2015), or converting high-dimensional data outright into functional data, where the latter has been referred to as Stringing (Wu & Müller 2010; Chen et al. 2011) and is based on a uni- or multi-dimensional scaling step to order predictors along locations on an interval or low-dimensional domain. The stringing method then assigns the value of the respective predictor to the location of the predictor on the interval, for all predictors. The distance of the predictor locations on the interval matches as closely as possible a distance measure between predictors that can be derived from correlations. Combining locations and predictor values and potentially also adding a smoothing step then converts the high-dimensional data for each subject or item to a random function. These functions can be summarized through their functional principal components, leading to an effective dimension reduction that is not based on sparsity and that works well for strongly correlated predictors.

Many recent developments in FDA have not been covered in this review. These include functional designs and domain selection problems and also dependent functional data such as functional time series, with many recent interesting developments, e.g. Panaretos & Tavakoli (2013). Another area that has gained recent interest are multivariate functional data. Similarly, in some longitudinal studies one observes for each subject repeatedly observed and therefore dependent functional data rather than scalars. There is also recently rising interest in spatially indexed functional data. These problems pose novel challenges for data analysis (Horvath & Kokoszka 2012).

While this review has focused on concepts and not on applications, as for other growing statistical areas, a driving force of recent developments in FDA has been the appearance of new types of data that require adequate methodology for their analysis. This is leading to the current development of “second generation” functional data that include more complex features than the first generation functional data that have been the emphasis of this review. Examples of recent applications include continuous tracking and monitoring of movement and health data, data that are recorded continuously over time by arrays of sensors, such as traffic flow data, continuously recorded climate and weather data, transcription factor count modeling along the genome, and the analysis of auction data, volatility and other financial data with functional methods.

Acknowledgments. We thank Pantelis Hadjipantelis for his help with creating the figures and an anonymous reviewer for very helpful comments that led to many improvements. JLW’s research was supported by NSF grants DMS-12-28369 and DMS-15-12975. HGM’s research was supported by NSF grants DMS-1228369 and DMS-1407852. JMC’s research was supported by Ministry of Science and Technology, Taiwan (MOST 104-2118-M-001-015-MY3).

LITERATURE CITED

Abraham C, Cornillon PA, Matzner-Lober E, Molinari N. 2003. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics* 30:581–595

Amini AA, Wainwright MJ. 2012. Sampled forms of functional pca in reproducing kernel hilbert spaces. *Annals of Statistics* 40:2483–2510

Angelini C, De Canditiis D, Pensky M. 2012. Clustering time-course microarray data using functional bayesian infinite mixture model. *Journal of Applied Statistics* 39:129–149

Araki Y, Konishi S, Kawano S, Matsui H. 2009. Functional logistic discrimination via regularized basis expansions. *Communications in Statistics-Theory and Methods* 38:2944–2957

Arribas-Gil A, Müller HG. 2014. Pairwise dynamic time warping for event data. *Computational Statistics and Data Analysis* 69:255–268

Ash RB, Gardner MF. 1975. Topics in Stochastic Processes. New York: Academic Press [Harcourt Brace Jovanovich Publishers]

Bali JL, Boente G, Tyler DE, Wang JL. 2011. Robust functional principal components: a projection-pursuit approach. *The Annals of Statistics* 39:2852–2882

Banfield JD, Raftery AE. 1993. Model-based gaussian and non-gaussian clustering. *Biometrics*

49:803–821. Times Cited: 727 1 739

Besse P, Ramsay JO. 1986. Principal components analysis of sampled functions. *Psychometrika* 51:285–311

Bickel P, Li B. 2007. Local polynomial regression on unknown manifolds. *Complex Datasets And Inverse Problems: Tomography, Networks And Beyond, ser. IMS Lecture Notes-Monograph Series.* 54:177–186

Bickel PJ, Rosenblatt M. 1973. On some global measures of the deviations of density function estimates. *The Annals of Statistics* 1:1071–1095

Boente G, Fraiman R. 2000. Kernel-based functional principal components. *Statistics & probability letters* 48:335–345

Boente G, Rodriguez D, Sued M. 2011. In *Recent advances in functional data analysis and related topics*, ed. F Ferraty, Contributions to Statistics. Springer, 49–53

Boente G, Salibián-Barrera M. 2014. S-estimators for functional principal component analysis. *Journal of the American Statistical Association* :Accepted

Bosq D. 2000. Linear processes in function spaces: theory and applications, vol. 149. Springer

Brumback B, Rice J. 1998. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 93:961–976

Cai T, Hall P. 2006. Prediction in functional linear regression. *Annals of Statistics* 34:2159–2179

Cai TT, Yuan M. 2011. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics* 39:2330–2355

Cao G, Yang L, Todem D. 2012. Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics* 24:359–377

Cardot H. 2000. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics* 12:503–538

Cardot H. 2007. Conditional functional principal components analysis. *Scandinavian Journal of Statistics* 34:317–335

Cardot H, Ferraty F, Sarda P. 1999. Functional linear model. *Statistics & Probability Letters* 45:11–22

Cardot H, Ferraty F, Sarda P. 2003. Spline estimators for the functional linear model. *Statistica Sinica* 13:571–592

Cardot H, Sarda P. 2005. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* 92:24–41

Cardot H, Sarda P. 2006. In *The Art of Semiparametrics*. Springer, 49–66

Carroll RJ, Maity A, Mammen E, Yu K. 2009. Nonparametric additive regression for repeatedly measured data. *Biometrika* 96:383–398

Castro PE, Lawton WH, Sylvestre EA. 1986. Principal modes of variation for processes with continuous sample curves. *Technometrics* 28:329–337

Chang C, Chen Y, Ogden RT. 2014. Functional data classification: a wavelet approach. *Computational Statistics*

Chen D, Hall P, Müller HG. 2011. Single and multiple index functional regression models with nonparametric link. *Annals of Statistics* 39:1720–1747

Chen D, Müller HG. 2012. Nonlinear manifold representations for functional data. *Annals of Statistics* 40:1–29

Chen K, Chen K, Müller HG, Wang J. 2011. Stringing high-dimensional data for functional analysis. *Journal of the American Statistical Association* 106:275–284

Chiou JM. 2012. Dynamical functional prediction and classification, with application to traffic flow prediction. *Annals of Applied Statistics* 6:1588–1614

Chiou JM, Li PL. 2007. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B* 69:679–699

Chiou JM, Li PL. 2008. Correlation-based functional clustering via subspace projection. *Journal of the American Statistical Association* 103:1684–1692

Chiou JM, Müller HG. 2007. Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis* 51:4849–4863

Chiou JM, Müller HG. 2009. Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association* 104:572–585

Chiou JM, Müller HG, Wang JL. 2003. Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65:405–423

Coffey N, Hinde J, Holian E. 2014. Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics & Data Analysis* 71:14–29

Conway JB. 1994. A course in functional analysis. Springer, 2nd ed.

Cook RD, Forzani L, Yao AF. 2010. Necessary and sufficient conditions for consistency of a method for smoothed functional inverse regression. *Statistica Sinica* 20:235–238

Crambes C, Delsol L, Laksaci A. 2008. Robust nonparametric estimation for functional data. *Journal of Nonparametric Statistics* 20:573–598

Şentürk D, Müller HG. 2005. Covariate adjusted correlation analysis via varying coefficient models. *Scandinavian Journal of Statistics* 32:365–383

Şentürk D, Müller HG. 2008. Generalized varying coefficient models for longitudinal data. *Biometrika* 95:653–666

Cuevas A, Febrero M, Fraiman R. 2004. An anova test for functional data. *Computational Statistics & Data Analysis* 47:111–122

Dauxois J, Pousse A, 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

de Boor C. 2001. A practical guide to splines, vol. 27. Springer Verlag

Degras D. 2008. Asymptotics for the nonparametric estimation of the mean function of a random process. *Statistics & Probability Letters* 78:2976–2980

Degras DA. 2011. Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica* 21:1735–1765

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

Delaigle A, Hall P. 2012. Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B* 74:267–286

Delaigle A, Hall P. 2013. Classification using censored functional data. *Journal of the American Statistical Association* 108:1269–1283

Donoho DL, Grimes C. 2005. Image manifolds which are isometric to euclidean space. *Journal of Mathematical Imaging and Vision* 23:5–24

Dou WW, Pollard D, Zhou HH. 2012. Estimation in functional regression for general exponential families. *Annals of Statistics* 40:2421–2451

Duan N, Li KC. 1991. Slicing regression: a link-free regression method. *The Annals of Statistics* 19:505–530

Dubin JA, Müller HG. 2005. Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association* 100:872–881

Eggermont PPB, Eubank RL, LaRiccia VN. 2010. Convergence rates for smoothing spline estimators in varying coefficient models. *Journal of Statistical Planning and Inference* 140:369–381

Eubank RL. 1999. Nonparametric regression and spline smoothing. CRC, 2nd ed.

Eubank RL, Hsing T. 2008. Canonical correlation for stochastic processes. *Stoch. Processes Appl.* 118:1634–1661

Fan J, Gijbels I. 1996. Local polynomial modelling and its applications. Chapman and Hall/CRC

Fan J, Lin SK. 1998. Test of significance when data are curves. *Journal of the American Statistical Association* 93:1007–1021

Fan J, Zhang W. 1999. Statistical estimation in varying coefficient models. *The Annals of Statistics* 27:1491–1518

Fan J, Zhang W. 2008. Statistical methods with varying coefficient models. *Statistics and Its Interface* 1:179–195

Fan Y, Foutz N, James GM, Jank W, et al. 2014. Functional response additive model estimation with online virtual stock markets. *Annals of Applied Statistics* 8:2435–2460

Ferraty F, Hall P, Vieu P. 2010. Most-predictive design points for functional data predictors. *Biometrika* 97:807–824

Ferraty F, Vieu P. 2003. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis* 44:161–173

Ferraty F, Vieu P. 2006. Nonparametric Functional Data Analysis. New York: Springer, New York

Ferré L, Yao AF. 2003. Functional sliced inverse regression analysis. *Statistics* 37:475–488

Ferré L, Yao AF. 2005. Smoothed functional inverse regression. *Statistica Sinica* 15:665–683

Garcia-Escudero LA, Gordaliza A. 2005. A proposal for robust curve clustering. *Journal of Classification* 22:185–201

Gasser T, Kneip A. 1995. Searching for structure in curve samples. *Journal of the American Statistical Association* 90:1179–1188

Gasser T, Müller HG, Köhler W, Molinari L, Prader A. 1984. Nonparametric regression analysis of growth curves. *Annals of Statistics* 12:210–229

Gervini D. 2008. Robust functional estimation using the median and spherical principal components. *Biometrika* 95:587–600

Gervini D. 2015. Warped functional regression. *Biometrika* :000–000

Giacofci M, Lambert-Lacroix S, Marot G, Picard F. 2013. Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics* 69:31–40

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

Grenander U. 1981. Abstract inference. Wiley New York

Hadjipantelis PZ, Aston JA, Müller HG, Evans JP. 2015. Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin chinese. *Journal of the American Statistical Association* :(in press)

Hall P, Horowitz JL. 2007. Methodology and convergence rates for functional linear regression. *Annals of Statistics* 35:70–91

Hall P, 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, Müller HG, Wang JL. 2006. Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics* 34:1493–1517

Hall P, Müller HG, Yao F. 2009. Estimation of functional derivatives. *Annals of Statistics* 37:3307–3329

Hall P, Poskitt DS, Presnell B. 2001. A functional dataanalytic approach to signal discrimination. *Technometrics* 43:1–9

Hall P, Van Keilegom I. 2007. Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica* 17:1511

Hastie T, Tibshirani R. 1986. Generalized additive models. *Statistical Science* 1:297–310

He G, Müller HG, Wang JL. 2000. In *Asymptotics in statistics and probability*, ed. ML Puri. VSP International Science Publishers

He G, Müller HG, Wang JL. 2003. Functional canonical analysis for square integrable stochastic processes. *Journal of Multivariate Analysis* 85:54–77

He G, Müller HG, Wang JL, Yang W. 2010. Functional linear regression via canonical analysis. *Bernoulli* 16:705–729

Heckman NE. 1986. Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society. Series B (Methodological)* :244–248

Heinzl F, Tutz G. 2013. Clustering in linear mixed models with approximate dirichlet process mixtures using em algorithm. *Statistical Modelling* 13:41–67

Heinzl F, Tutz G. 2014. Clustering in linear-mixed models with a group fused lasso penalty. *Biometrical Journal* 56:44–68

Hilgert N, Mas A, Verzelen N. 2013. Minimax adaptive tests for the functional linear model. *Annals of Statistics* 41:838–869

Hoover D, Rice J, Wu C, Yang L. 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85:809–822

Horvath L, Kokoszka P. 2012. Inference for functional data with applications. New York: Springer

Horváth L, Reeder R. 2013. A test of significance in functional quadratic regression. *Bernoulli* 19:2120–2151

Hsing T, Eubank R. 2015. Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons

Hu Z, Wang N, Carroll RJ. 2004. Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika* 91:251–262

Huang J, Wu C, Zhou L. 2002. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* 89:111–128

Huang J, Wu C, Zhou L. 2004. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 14:763–788

Hyndman RJ, Shang HL. 2010. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics* 19:29–45

Jacques J, Preda C. 2013. Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing* 112:164–171

Jacques J, Preda C. 2014. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* 71:92–106

James G, Hastie T, Sugar C. 2000. Principal component models for sparse functional data. *Biometrika* 87:587–602

James GM. 2002. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B* 64:411–432

James GM, Hastie TJ. 2001. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 63:533–550

James GM, Silverman BW. 2005. Functional adaptive model estimation. *Journal of the American Statistical Association* 100:565–576

James GM, Sugar CA. 2003. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98:397–408

Jiang C, Aston JA, Wang JL. 2009. Smoothing dynamic positron emission tomography time courses using functional principal components. *NeuroImage* 47:184–193 (PMC2715874)

Jiang C, Wang JL. 2010. Covariate adjusted functional principal components analysis for longitudinal data. *The Annals of Statistics* 38:1194–1226

Jiang C, Wang JL. 2011. Functional single index models for longitudinal data. *The Annals of Statistics* 39:362–388

Jiang C, Yu W, Wang JL. 2014. Inverse regression for longitudinal data. *The Annals of Statistics* 42:563—591

Jolliffe I. 2002. Principal component analysis. Springer, 2nd ed.

Jones MC, Rice JA. 1992. Displaying the important features of large collections of similar curves. *The American Statistician* 46:140–145

Karhunen K. 1946. Zur Spektraltheorie stochastischer Prozesse. *Annales Academiae Scientiarum Fennicae. Series A. I, Mathematica* 1946:7

Kato T. 1980. Perturbation theory for linear operators. Springer, 2nd ed.

Kayano M, Dozono K, Konishi S. 2010. Functional cluster analysis via orthonormalized gaussian basis expansions and its application. *Journal of Classification* 27:211–230

Kirkpatrick M, Heckman N. 1989. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* 27:429–450

Kneip A, Gasser T. 1988. Convergence and consistency results for self-modeling nonlinear regression. *Annals of Statistics* 16:82–112

Kneip A, Gasser T. 1992. Statistical tools to analyze data representing a sample of curves. *Annals of Statistics* 20:1266–1305

Kneip A, Ramsay JO. 2008. Combining registration and fitting for functional models. *Journal of the American Statistical Association* 103:1155–1165

Kneip A, Sarda P. 2011. Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics* 39:2410–2447

Kneip A, Utikal KJ. 2001. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* 96:519–542

Kong D, Xue K, Yao F, Zhang HH. 2015. Partially functional linear regression in high dimensions. *Biometrika* in press

Kraus D, Panaretos VM. 2012. Dispersion operators and resistant second-order functional data analysis. *Biometrika* 99:813–832

Lai RCS, Huang HC, Lee TCM. 2012. Fixed and random effects selection in nonparametric additive mixed models. *Electronic Journal of Statistics* 6:810–842

Lawton WH, Sylvestre EA. 1971. Self modeling curve resolution. *Technometrics* 13:617–633

Leng X, Müller HG. 2006. Time ordering of gene co-expression. *Biostatistics* 7:569–584

Leurgans SE, Moyeed RA, Silverman BW. 1993. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)* 55:725–740

Li KC. 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* 86:316–327

Li PL, Chiou JM. 2011. Identifying cluster number for subspace projected functional data clustering. *Computational Statistics & Data Analysis* 55:2090–2103

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

Lin X, Zhang D. 1999. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Serial B (Methodological)* 61:381–400

Little RJ, Rubin DB. 2014. Statistical analysis with missing data. John Wiley & Sons

Liu B, Müller HG. 2009. Estimating derivatives for samples of sparsely observed functions, with application to on-line auction dynamics. *Journal of the American Statistical Association* 104:704–714

Liu X, Müller HG. 2004. Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association* 99:687–699

Loève M. 1946. Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique* 84:159–162

Ma S, Yang L, Carroll RJ. 2012. A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica* 22:95

Malfait N, Ramsay JO. 2003. The historical functional linear model. *Canadian Journal of Statistics* 31:115–128

Matsui H, Araki T, Konishi S. 2011. Multiclass functional discriminant analysis and its application to gesture recognition. *Journal of Classification* 28:227–243

McCullagh P, Nelder JA. 1983. Generalized linear models. Monographs on Statistics and Applied Probability. London: Chapman & Hall

McLean MW, Hooker G, Staicu AM, Scheipl F, Ruppert D. 2014. Functional generalized additive models. *Journal of Computational and Graphical Statistics* 23:249–269

Morris JS. 2015. Functional regression. *Annual Review of Statistics and Its Application* 2:Online

Müller HG. 2005. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* 32:223–240

Müller HG. 2008. In *Longitudinal Data Analysis (Handbooks of Modern Statistical Methods)*, eds.

G Fitzmaurice, M Davidian, G Verbeke, G Molenberghs. New York: Chapman & Hall/CRC, 223–252

Müller HG. 2011. In *International Encyclopedia of Statistical Science*, ed. M Lovric. Springer, Heidelberg, 554–555. (Extended version available in StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies, id 242).

Müller HG, Carey JR, Wu D, Liedo P, Vaupel JW. 2001. Reproductive potential predicts longevity of female Mediterranean fruit flies. *Proceedings of the Royal Society of London - Series B : Biological Science* 268:445–450

Müller HG, Stadtmüller U. 2005. Generalized functional linear models. *Annals of Statistics* 33:774–805

Müller HG, Wu S, Diamantidis AD, Papadopoulos NT, Carey JR. 2009. Reproduction is adapted to survival characteristics across geographically isolated medfly populations. *Proceedings of the Royal Society of London - Series B : Biological Science* 276:4409–4416

Müller HG, Wu Y, Yao F. 2013. Continuously additive models for nonlinear functional regression. *Biometrika* 100:607–622

Müller HG, Yao F. 2008. Functional additive models. *Journal of the American Statistical Association* 103:1534–1544

Müller HG, Yao F. 2010a. Additive modeling of functional gradients. *Biometrika*

Müller HG, Yao F. 2010b. Empirical dynamics for longitudinal data. *Annals of Statistics* 38:3458–3486

Opgen-Rhein R, Strimmer K. 2006. Inferring gene dependency networks from genomic longitudinal data: A functional data approach. *REVSTAT - Statistical Journal* 4:53–65

Panaretos VM, Kraus D, Maddocks JH. 2010. Second-order comparison of gaussian random functions and the geometry of dna minicircles. *Journal of the American Statistical Association* 105:670–682

Panaretos VM, Tavakoli S. 2013. Fourier analysis of stationary time series in function space. *Annals of Statistics* 41:568–603

Paul D, Peng J. 2009. Consistency of restricted maximum likelihood estimators of principal components. *The Annals of Statistics* 37:1229–1271

Peng J, Müller HG. 2008. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Annals of Applied Statistics* 2:1056–1077

Peng J, Paul D, Müller HG. 2014. Time-warped growth processes, with applications to the modeling of boom–bust cycles in house prices. *Annals of Applied Statistics* 8:1561–1582

Petrone S, Guindani M, Gelfand AE. 2009. Hybrid dirichlet mixture models for functional data. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 71:755–782

Pezzulli S, Silverman B. 1993. Some properties of smoothed principal components analysis for functional data. *Computational Statistics* 8:1–1

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Ramsay JO. 1982. When the data are functions. *Psychometrika* 47:379–396

Ramsay JO, Dalzell C. 1991. Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* :539–572

Ramsay JO, Hooker G, Graves S. 2009. Functional data analysis with r and matlab. Springer

Ramsay JO, Li X. 1998. Curve registration. *Journal of the Royal Statistical Society: Series B* 60:351–363

Ramsay JO, Silverman BW. 2002. Applied functional data analysis: methods and case studies, vol. 77. Springer

Ramsay JO, Silverman BW. 2005. Functional Data Analysis. Springer Series in Statistics. New York: Springer, 2nd ed.

Rao CR. 1958. Some statistical methods for comparison of growth curves. *Biometrics* 14:1–17

Rice J, Silverman B. 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)* 53:233–243

Rice JA. 2004. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica* 14:631–647

Rice JA, Wu CO. 2001. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* 57:253–259

Rincon M, Ruiz-Medina MD. 2012. Wavelet-rkhs-based functional statistical classification. *Advances in Data Analysis and Classification* 6:201–217

Rodriguez A, Dunson DB, Gelfand AE. 2009. Bayesian nonparametric functional data analysis through density estimation. *Biometrika* 96:149–162

Roweis ST, Saul LK. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290

Sakoe H, Chiba S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26:43–49

Serban N, Wasserman L. 2005. Cats: Clustering after transformation and smoothing. *Journal of the American Statistical Association* 100:990–999. Times Cited: 31 0 31

Shi M, Weiss RE, Taylor JM. 1996. An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics* 45:151–163

Silverman BW. 1995. Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society: Series B* 57:673–689

Silverman BW. 1996. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* 24:1–24

Sood A, James G, Tellis GJ. 2009. Functional regression: a new model for predicting market penetration of new products of new products. *Marketing Science* 28:36–51. (to appear)

Speckman P. 1988. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* 50:413–436

Staniswalis JG, Lee JJ. 1998. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 93:1403–1418

Stone CJ. 1985. Additive regression and other nonparametric models. *The Annals of Statistics* 13:689–705

Sun Y, Genton MG. 2011. Functional boxplots. *Journal of Computational and Graphical Statistics* 20:316–334

Tang R, Müller HG. 2008. Pairwise curve synchronization for functional data. *Biometrika* 95:875–889

Tenenbaum JB, De Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–2323

Tucker JD, Wu W, Srivastava A. 2013. Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis* 61:50–66

Verzelen N, Tao W, Müller HG. 2012. Inferring stochastic dynamics from functional data. *Biometrika* 99:533–550

Wand MP, Jones CM. 1995. Kernel smoothing. Chapman and Hall/CRC

Wang J, Yang L. 2009. Polynomial spline confidence bands for regression curves. *Statistica Sinica* 19:325–342

Wang K, Gasser T. 1997. Alignment of curves by dynamic time warping. *Annals of Statistics* 25:1251–1276

Wang S, Qian L, Carroll RJ. 2010. Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika* 97:79–93

Wang XH, Ray S, Mallick BK. 2007. Bayesian curve classification using wavelets. *Journal of the American Statistical Association* 102:962–973

Wu CO, Chiang CT. 2000. Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica* 10:433–456

Wu CO, Chiang CT, Hoover DR. 1998. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association* 93:1388–1402

Wu H, Zhang JT. 2006. Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches. Wiley-Interscience

Wu P, Müller HG. 2010. Functional embedding for the classification of gene expression profiles. *Bioinformatics* 26:509–517

Wu W, Srivastava A. 2014. Analysis of spike train data: Alignment and comparisons using the extended fisher-rao metric. *Electronic Journal of Statistics* 8:1776–1785

Xia Y, Tong H, Li W, Zhu LX. 2002. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64:363–410

Yang W, Müller HG, Stadtmüller U. 2011. Functional singular component analysis. *Journal of the Royal Statistical Society: Series B* 73:303–324

Yao F, Lee T. 2006. Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68:3–25

Yao F, Müller HG. 2010. Functional quadratic regression. *Biometrika* 97:49–64

Yao F, Müller HG, Wang JL. 2005a. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100:577–590

Yao F, Müller HG, Wang JL. 2005b. Functional linear regression analysis for longitudinal data. *The Annals of Statistics* 33:2873–2903

You J, Zhou H. 2007. Two-stage efficient estimation of longitudinal nonparametric additive models. *Statistics & Probability Letters* 77:1666–1675

Zhang JT. 2013. Analysis of variance for functional data. Hall/Crc Monographs on Statistics & Applied Probability. Taylor & Francis Group

Zhang X, Park BU, Wang JL. 2013. Time-varying additive models for longitudinal data. *Journal of the American Statistical Association* 108:983–998

Zhang X, Wang JL. 2014. From sparse to dense functional data and beyond. Tech. rep., University of California, Davis

Zhang X, Wang JL. 2015. Varying-coefficient additive models for functional data. *Biometrika* :In Press

Zhang Z, Müller HG. 2011. Functional density synchronization. *Computational Statistics and Data Analysis* 55:2234–2249

Zhao X, Marron JS, Wells MT. 2004. The functional data analysis view of longitudinal data. *Statistica Sinica* 14:789–808

Zhu H, Fan J, Kong L. 2014. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association* 109:1084–1098

Zhu HX, Brown PJ, Morris JS. 2012. Robust classification of functional and quantitative image data using functional mixed models. *Biometrics* 68:1260–1268

Zhu HX, Vannucci M, Cox DD. 2010. A bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* 66:463–473