Functional Modeling of Longitudinal Data

Hans-Georg Müller

Department of Statistics, University of California,

One Shields Ave., Davis, CA 95616, U.S.A.

mueller@wald.ucdavis.edu

July 29, 2006

SUMMARY

Functional data analysis provides an inherently nonparametric approach for the analysis of data which consist of samples of time courses or random trajectories. It is a relatively young field aiming at modeling and data exploration under very flexible model assumptions with no or few parametric components. Basic tools of functional data analysis are smoothing, functional principal components, functional linear models and time-warping. Warping or curve registration aims at adjusting for random time distortions. While in the usual functional data analysis paradigm the sample functions were considered as continuously observed, in longitudinal data analysis one mostly deals with sparsely and irregularly observed data that also are corrupted with noise. Adjustments of functional data analysis techniques which take these particular features into account are needed to use them to advantage for longitudinal data. We review some techniques that have been recently proposed to connect functional data analysis methodology with longitudinal data. The extension of functional data analysis towards longitudinal data is a fairly recent undertaking that presents a promising avenue for future research. This article provides a review of some of the recent developments.

Key words: Conditioning, Covariance function, Eigenfunction, Functional Principal Component, Functional Regression, Karhunen-Loève expansion, Kernel method, Local linear fitting, Prediction, Registration, Regression parameter function, Smoothing, Trajectories, Sparse Data, Stochastic Process, Warping.

1 Introduction

Longitudinal studies are characterized by data records containing repeated measurements per subject, measured at various points on a suitable time axis. The aim is often to study change over time or time-dynamics of biological phenomena such as growth, physiology, pathophysiology and pathogenesis. One is also interested in relating these time-dynamics to certain predictors or responses. The classical analysis of longitudinal studies is based on parametric models which often contain random effects such as the Generalized Linear Mixed Model (GLMM) or are marginal methods such as Generalized Estimating Equations (GEE). The relationships between the subject-specific random effects models and the marginal population-average models such as GEE are quite complex (see, e.g., Zeger, Liang and Albert (1988), Heagerty (1999), Heagerty and Zeger (2000)). To a large extent, this non-compatibility of various approaches is due to the parametric assumptions that are made in these models. These include the

assumption of a parametric trend (linear or quadratic in the simplest cases) over time and of parametric link functions. Specific common additional assumptions are normality of the random effects in a GLMM and a specific covariance structure ("working correlation") in a GEE. Introducing nonparametric components (nonparametric link and nonparametric covariance structure) can ameliorate the difficulties of relating various longitudinal models to each other, as it increases the inherent flexibility of the resulting longitudinal models substantially (compare the Estimated Estimating Equations approach in Chiou and Müller, 2005).

Taking the idea of modeling with nonparametric components one step further, the Functional Data Analysis (FDA) approach to longitudinal data provides an alternative nonparametric method for the modeling of individual trajectories. The underlying idea is to view observed longitudinal trajectories as a sample of random functions, which are not parametrically specified. The observed measurements for an individual then correspond to the values of the random trajectory, corrupted by some measurement error. A primary objective is to reduce the high dimension of the trajectories – considered to be elements of an infinite-dimensional function space – to finite dimension. One goal is to predict individual trajectories from the measurements made for a subject, borrowing strength from the entire sample of subjects. The necessary dimension reduction or regularization step can be implemented in various ways. For the analysis of longitudinal data, with its typically sparse and irregular measurements per subject, the method of Functional Principal Component Analysis (FPCA) has been recently proposed (Yao, Müller and Wang, 2005ab), extending previous work by James (2002). Other regularization methods that have proven useful in FDA include smoothing splines (Ke and Wang 2001), B-splines (Rice and Wu 2000) or P-splines (Yao and Lee, 2006).

The classical theory and applications of FDA have been developed for densely sampled or fully observed trajectories that in addition are sampled without noise. This setting is not conducive to applications in longitudinal studies, due to the common occurrence of irregular and sparse measurement times, often due to missing data. Excellent overviews on FDA for densely sampled data or fully observed trajectories can be found in the two recent books by Ramsay and Silverman (2002, 2005). Early approaches were based primarily on smoothing techniques and landmarks (e.g., Gasser et al., 1984, 1985). The connections between FDA with longitudinal data analysis have been revisited more recently, compare Rice (2004), Zhao, Marron and Wells (2004) and Müller (2005). Of interest is also a discussion that was held in 2004 at a conference dedicated to exploring these connections (Marron et al. 2004). While a number of practical procedures and also theoretical results are available, the use of FDA methodology for the analysis of longitudinal data is far from being established practice. This is an area of ongoing research.

Even the estimation of a mean trajectory is non-trivial: Dependency of the repeated measurements coming from the same subject may be taken into account (Lin et al. 2004, Wang 2003) to improve efficiency of this estimation step. Another problem with major practical impact occurs for longitudinal studies, where individually varying time scales matter. In such situations, warping approaches may be needed, as discussed in Section 3 below.

We focus in the following on an approach of applying FDA to longitudinal data that is based on FPCA and thus allows for subject-specific models that include random effects and which are entirely data-adaptive (section 5). Our focus is less on marginal population-average modeling, although we discuss below the difficulties that are caused for marginal modeling in the presence of warping. Auxiliary quantities of interest include estimates of the underlying population-average mean function and of the covariance surface describing the dependency structure of the repeated measurements. These steps require smoothing methods, briefly reviewed in the next section.

From the covariance surface, one estimates the eigenfunctions of the underlying stochastic process that is assumed to generate the individual random trajectories. We do not assume stationarity. Individual trajectories are represented by their first few functional principal component (FPC) scores. These scores play the role of random effects. Thus, functional data are reduced to a vector of scores. These scores may subsequently be entered into further statistical analysis, either serving as predictors or as responses in various statistical models, including functional regression models (section 6).

2 Basics of Functional Data Analysis

Functional data consist of a sample of random curves which are typically viewed as i.i.d. realizations of an underlying stochastic process. Per subject or experimental unit, one samples one or several functions $Y(t)$, $t \in \mathcal{T}$, where $\mathcal{T}$ is a suitable domain, usually an interval. A common assumption is that trajectories are square integrable and smooth, say twice differentiable. A major difference between functional data and multivariate data is that in the case of functional data, order and neighborhood relations are well defined and relevant, i.e., one has a topology on the domain on which the trajectories are defined. In contrast, for multivariate data, irrespective of dimension, no meaningful topology on the domain exists. This is illustrated by the fact that one can re-order the components of a multivariate data vector and arrive at exactly the same statistical analysis as for the data vector arranged in the original order. For functional data, the situation is entirely different.

Goals for FDA include the construction of meaningful models for basic data descriptive measures such as a mean trajectory. If one was given a sample of entirely observed trajectories $Y_i(t)$, $i = 1, \ldots, N$, for

N subjects, a mean trajectory could be simply defined as sample average, $\bar{Y}(t) = \frac{1}{N}\sum_{i=1}^{N} Y_i(t)$, $t \in T$. However, this relatively straightforward situation is rather the exception than the norm, as we face the following difficulties: The trajectories may be sampled at sparsely distributed times, with timings varying from subject to subject; the measurements may be corrupted by noise and are dependent within the same subject; and time-warping may be present, a complication that is typical for some longitudinal data such as growth curves and is discussed further in Section 4. So what constitutes a reasonable population mean function is much less straightforward in the FDA setting than it is in the multivariate case. Further notions of interest which require special attention include measures of variance, covariance and correlation between curves.

Measures of correlation are of interest for studies in which several trajectories per subject are observed. An initial idea has been the extension of canonical correlation from the multivariate (Hotelling 1936) to the functional case. The resulting functional canonical correlation requires inversion of a compact linear operator which makes this an inverse problem. Such problems require regularization. Two main types of regularization have been used in FDA: Regularization by an additive penalty term, usually penalizing against non-smooth curve estimates, and used in combination with spline modeling; or truncation of a functional series expansion such as a Fourier series or wavelet expansion, at a finite number of terms, also known as thresholding. Both approaches depend on the choice of an appropriate regularization parameter. For functional canonical correlation, both regularization by a penalty (Leurgans, Moyeed and Silverman 1993) and by truncation (He, Müller and Wang 2004) have been proposed. One consistent finding is that functional canonical correlation is highly sensitive to the choice of the regularization parameter (size of penalty or truncation). Due to the difficulties in calibrating the regularization for functional canonical correlation, alternative notions of a functional dynamic correlation (Dubin and Müller, 2005; compare also Service, Rice and Chavez 1998, Heckman and Zamar 2000) have been studied.

Beyond functional correlation, the problem of relating several observed curves per subject to each other or to a scalar response leads to the problem of functional regression. Functional regression models come in various flavors: For a scalar response, one may consider one or several functional predictors. There are also situations in which the response is a function, combined with scalar or multivariate predictors. The most complex case involves the simultaneous presence of functional predictors and functional responses. These models will be discussed in Section 6. In functional regression, one can distinguish a classic FDA approach which requires the availability of fully observed noise-free individual trajectories and has been well investigated in recent years (Ramsay and Dalzell, 1991; Cardot et al. 2003) and a modified approach suitable for longitudinal data that is of more recent origin. Methods extending

density estimation and nonparametric regression to functional objects have also been developed in recent years (Ferraty and Vieu 2006); such developments face theoretical challenges and are the subject of ongoing research.

We take here functional data to mean one deals with a sample of curves, rather than with a single curve such as in dose-response analysis or in nonparametric regression function. The use of "functional data" is however not always that rigorous and often simply refers to the fact that a model contains a nonparametric curve as a component.

When faced with functional data, a useful first step is to simply plot the data. In situations characterized by reasonably dense sampling of measurements per subject, one may generate such plots by linearly interpolating the points corresponding to the repeated measurements made on the same subject ("spaghetti plot"). In other situations, when data are irregularly sampled or a derivative is required as in the as in the modeling of growth curves, it is best to conduct a preliminary smoothing or differentiation step before plotting the data (see Fig. 2 and 3, where in the left panel of Fig. 2 an initial kernel differentiation was implemented, while Fig. 3 shows spaghetti plots of unprocessed data). From such plots, one may discern a general trend in the data, changes in sampling frequencies (for example caused by drop-outs) and shapes of individual trajectories and their variation. Last not least one may identify subjects with outlying trajectories; these are candidates for removal before proceeding with the analysis.

FDA relies on and exploits smoothing due to the smooth topology in the domain that distinguishes FDA from multidimensional data analysis. Another characteristic feature of functional data is the presence of warping, i.e., the possibility that individual time scales are randomly distorted. Some basic smoothing ideas are discussed in the following subsection, while the next section is devoted to an introduction to warping (also known as curve registration or curve alignment).

3 Nonparametric Regression

3.1 Kernel smoothing

Smoothing methods for nonparametric regression are an important ingredient of FDA, as the key techniques exploit the continuity of the trajectories. We focus here on kernel-type smoothers that have proven useful, due to their straightforward interpretation and the large body of accumulated knowledge about their properties, especially their asymptotic behavior. Explicit representations in terms of weighted averages in the data, which are available for this class of smoothers, greatly facilitate the investigation of asymptotic properties and also of the FDA methods that utilize them. Excellent textbooks

and monographs on kernel-type smoothing procedures include Bowman and Azzalini (1997), Fan and Gijbels (1996), Silverman (1986) or Wand and Jones (1995). Other smoothing methods such as various types of splines can often be used equally well in nonparametric regression (Eubank 1999).

The goal of smoothing in the nonparametric regression setting is to estimate a smooth regression function or surface $g(u) = E(V|U = u)$, usually assumed to be twice continuously differentiable. For the random design case, this regression function is characterized by the joint distribution of vectors (U, V), while for fixed designs the predictor levels U_j, at which responses V_j are recorded, are assumed to be non-random (and usually assumed to be generated by a design density). The response V is univariate, while predictors U can be univariate or multivariate. Of interest for FDA applications are the cases $u \in \mathbf{R}$, the case of a one-dimensional regression function, and $u \in \mathbf{R}^2$, the case of a regression surface.

To define a kernel smoother for a one-dimensional predictor, given n data points $\{(U_j, V_j)_{1 \le j \le n}\}$, we need a bandwidth or window width h and a kernel function K. The bandwidth serves as smoothing parameter and determines the trade-off between variance and bias of the resulting nonparametric regression estimates. The kernel K typically is chosen as a smooth and symmetric density function; for some types of kernel estimators such as convolution estimators, negative valued kernels can be used to accelerate rates of convergence (Gasser, Müller and Mammitzsch 1985). Commonly used non-negative kernels are rectangular (box) kernels $K(x) = 1$, quadratic (Epanechnikov) kernels $K(x) = \frac{3}{4}(1 - x^2)$, which enjoy some optimality properties, and Gaussian kernels which correspond to the standard normal density.

A classic kernel smoothing method primarily aimed at regular designs U_j is the class of convolution kernel smoothers (Priestley and Chao 1972; Gasser and Müller 1984). The smoothing window for estimating at predictor level u is $[u - h, u + h]$ if a kernel function K with domain $[-1, 1]$ is used. Let $S_{(j)} = (U_{(j)} + U_{(j-1)})/2$, where $U_{(j)}$ is the j-th order statistic of the U_j, and let $V_{[j]}$ denotes the concomitant of $U_{(j)}$. Convolution type kernel estimators are defined as

$$\hat{g}_C(u) = \sum_{j=1}^{n} V_{[j]} \int_{S_{(j)}}^{S_{(j+1)}} \frac{1}{h} K(\frac{u - s}{h}) ds. \tag{1}$$

Near the endpoints of the regression function, specially constructed boundary kernels should be used to avoid boundary bias effects (e.g., Jones and Foster 1996; Müller 1991).

3.2 Extensions and local linear fitting

We note that these smoothers can be easily extended to the case of estimating derivatives (Gasser and Müller 1984). Convolution type smoothers have been applied extensively to conduct nonparametric

analysis of longitudinal growth studies (Gasser et al. 1984, Müller 1988). Growth studies belong to a class of longitudinal studies for which one has relatively dense measurement grids. In such situations one can smooth each trajectory individually, independent of the other observed trajectories. This is justified by postulating asymptotically ever denser designs where the number of measurements per subject n increases asymptotically within a fixed domain (also referred to as in-fill asymptotics). As $n \to \infty$, using appropriate kernels and bandwidth sequences, this approach leads to estimates of trajectories and derivatives with typical nonparametric rates of convergence of the order $n^{-(k-\nu)/(2k+1)}$. Here, ν is the order of derivative to be estimated and $k > \nu$ is the order of assumed smoothness of the trajectory (number of continuous derivatives). For an example of a sample of estimated first derivatives of growth data, see the left panel of Fig. 2.

The analyses of growth data with these smoothing methods demonstrated that nonparametric regression methods are essential tools to discern features of longitudinal time courses. An example is the detection of a pre-pubertal growth spurt which had been omitted from previously used parametric models. Once a longitudinal feature is not properly reflected in a parametric model, it can be very difficult to discover these features through a lack-of-fit analysis. A nonparametric approach should always be used concurrently with a parametric modeling approach in order to ensure against omitting important features. Nonparametric methods achieve this by being very flexible and by not reflecting preconceived notions about the shape of time courses. In the above mentioned analysis of growth studies, first and second derivatives where estimated for each individual separately to assess the dynamics of growth. For the practically important problem of bandwidth choice, one can use cross-validation (minimization of one-leave-out prediction error), generalized cross-validation (a faster approximation) or a variety of plug-in methods aiming at minimizing Mean Squared Error or Integrated Mean Squared Error.

Boundary adjustments are automatically included in local polynomial fitting, which is a great advantage. Local linear smoothers are particularly easy to use and have become the most popular kernel-based smoothing method. They are based on the very simple idea of localizing a simple linear regression from the entire data domain to local windows and have been around for a long time. Compared to convolution kernel estimators, this method has better conditional variance properties in random designs. A theoretical problem is that the unconditional variance of local linear estimators is unbounded, therefore mean squared error does not exist, in contrast to the convolution methods where it is always bounded. Practically, this is reflected by problems caused by occasional gaps in the designs, i.e., for a random design the probability that not enough data fall into at least one smoothing window is not negligible (see Seifert and Gasser 1996, 2000, for further discussion of these issues and improved local linear estimation).

The local linear kernel smoother (Fan and Gijbels, 1996) is obtained via the minimizers $\hat{a}_0, \hat{a}_1$ of

$$\sum_{j=1}^{M} K\left(\frac{u - U_j}{b}\right) [V_i - a_0 - a_1(u - U_i)]^2 , \tag{2}$$

setting $\hat{g}_L(u) = \hat{a}_0$. The older kernel methods of Nadaraya (1964) and Watson (1964) correspond to the less flexible special case of local linear fitting where one fits local constants, which leads to somewhat awkward bias behavior.

For two-dimensional smoothing we aim at the regression function $g(u_1, u_2) = E(V|U_1 = u_1, U_2 = u_2)$. Locally weighted least squares then provide a criterion for fitting local planes to the data $\{(U_{j1}, U_{j2}, V_j)_{j=1,\ldots,M}\}$, leading to the surface estimate $\hat{g}(u_1, u_2) = \hat{a}_0$, where $(\hat{a}_0, \hat{a}_1, \hat{a}_2)$ are the minimizers of the locally weighted sum of squares

$$\sum_{j=1}^{M} K\left(\frac{u_1 - U_{j1}}{h_1}, \frac{u_2 - U_{i2}}{h_2}\right) [V_j - (a_0 + a_1(U_{j1} - u_1) + a_2(U_{j2} - u_2)]^2 \ . \tag{3}$$

Here, $K(\cdot, \cdot) \geq 0$ is a two-dimensional kernel function. It can be chosen as a product of two one-dimensional kernel functions. Two bandwidths h_1, h_2 are needed, for simplicity they are often chosen to be the same. We note that explicit formulas for these smoothers can be easily obtained (see, e.g., formulas (2.5) in Hall, Müller and Wang, 2006).

4 Time Warping and Curve Synchronization

4.1 Overview

Time-warping has been studied primarily for densely sampled data, such as longitudinal growth studies, but is of potential relevance for many longitudinal studies. Warping is also referred to as time synchronization, curve registration or curve alignment (Gasser and Kneip 1995, Ramsay and Li 1998, Rønn 2001, Liu and Müller 2004, Gervini and Gasser, 2004). In a warping model one assumes that the time axis is individually distorted, for example by a random time transformation function that is monotone increasing and keeps beginning and end point of the domain as fixed. The motivation for considering warping in biomedical applications is that individuals may progress through time at their own individual pace, referred to as biological time or eigenzeit (Capra and Müller 1997) which may differ from clock time. A typical example is human growth where the various growth spurts (the pubertal spurt but also the so-called mid-growth spurt that has been rediscovered using nonparametric smoothing methods – see Gasser et al. 1984) occur at different ages for different individuals. In such a situation, a cross-sectional average growth curve will often not be meaningful. The reason is that it

will not resemble any individual trajectory closely and therefore gives a wrong impression about the dynamics of growth.

An example is shown in Fig. 1; the left panel displays a sample of growth velocities from 54 girls of the Berkeley Growth Study. The right panel features a comparison the cross-sectional mean growth curve with various warped means. Among these, the landmark method, pioneered in Kneip and Gasser (1992) is known to work very well for these data; the Procrustes method (Ramsay and Li 1998) is an iterative procedure, warping curves at each step to match the current cross-sectional mean; and area-under-the-curve registration (introduced in Liu and Müller 2004) synchronizes time points associated with the same relative area under the curve between the left endpoint of the domain and this time point. The cross-sectional mean is found to underestimate the size of the pubertal growth spurt and also produces a biased location, and similar distortions for the midgrowth spurt, the smaller growth growth spurt at around 5 years.

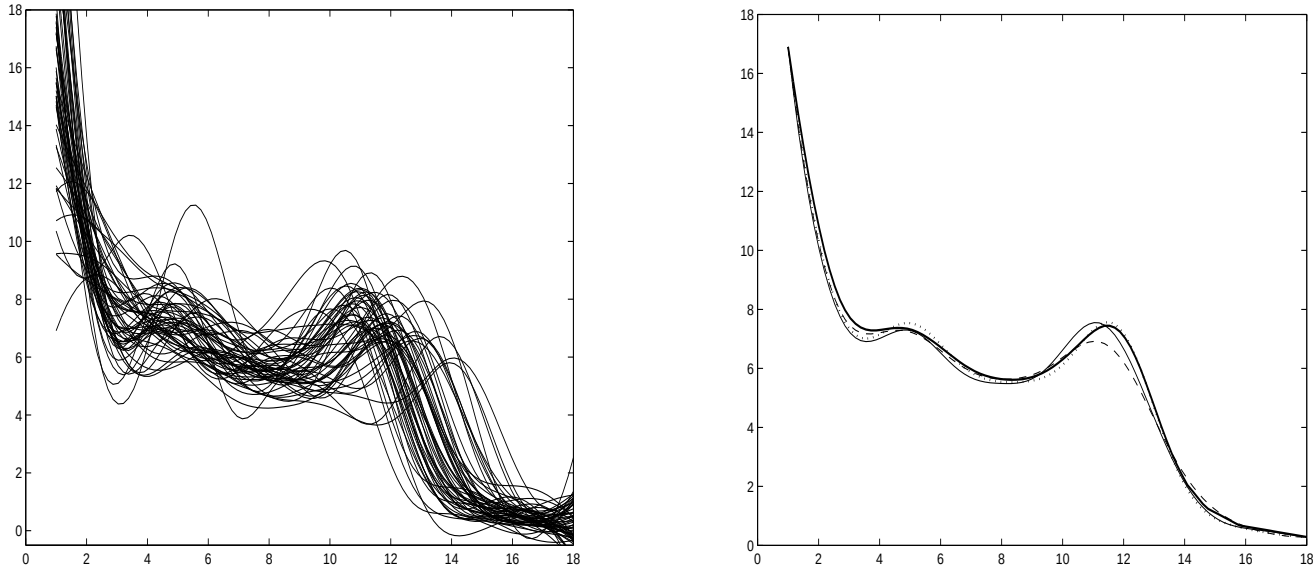

Figure 1: Time-warping of growth curves. Left panel: Sample of estimated growth velocities (first derivatives) of 54 girls from the Berkeley Growth Study. Right panel: Comparison of different registration procedures, including functional convex mean using area-under-the-curve registration (solid bold), continuous monotone registration (so-called Procrustes method, solid), landmark registration (dotted), and cross-sectional average (dashed). In both panels, $x-$ axis is age in years and $y-$ axis is velocity in cm/yr. Figure reproduced from Liu and Müller (2004). Functional convex averaging and synchronization for time-warped random curves. *J. Amer. Statist. Assoc.* **99**, 687-699.

The conclusion is that even a simple notion such as a mean needs to be carefully considered in the

presence of warping. A simulated example demonstrating the distorting effects of warping in FDA is shown in Fig. 2. Here the individual trajectories have been generated as bimodal curves but the cross-sectional mean does not reflect this shape at all. Therefore, modifications aiming at time-synchronization are needed to arrive at a representative mean when warping is present.

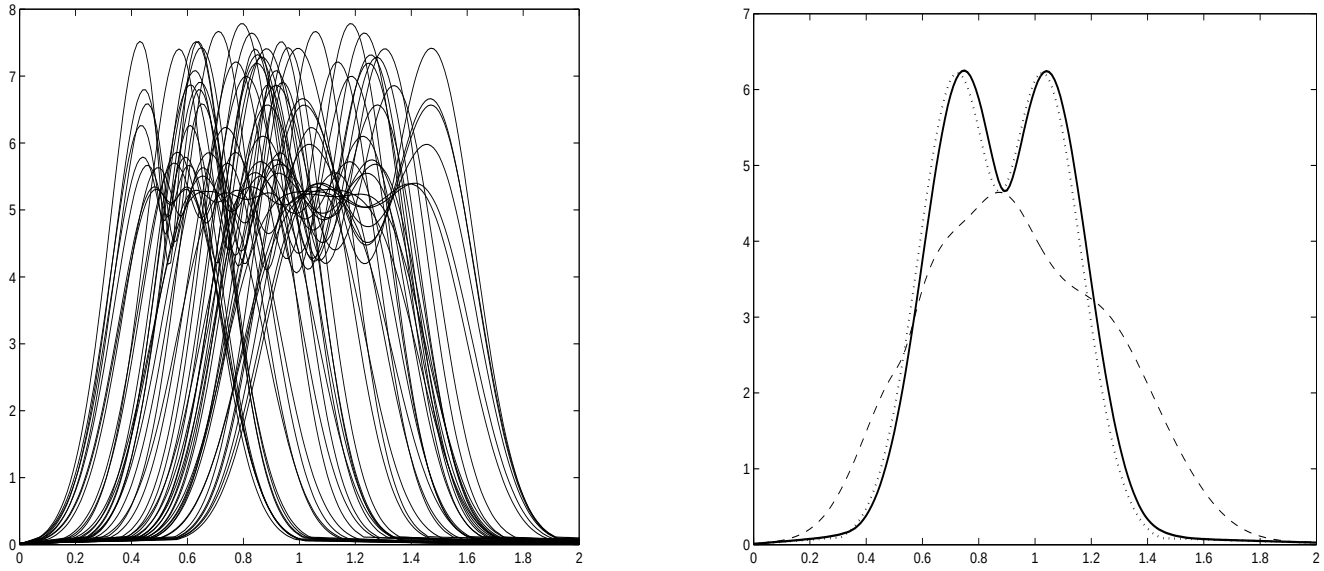

Figure 2: The effect of warping. Left panel: Sample of simulated bimodal trajectories. Right panel: Target is the true bimodal functional warped mean (solid). Estimated means are the naive cross-sectional mean computed from the sample curves (dashed) and the warped functional mean computed from the sample curves, using area-under-the-curve registration (dotted). Figure reproduced from Liu and Müller (2004). Functional convex averaging and synchronization for time-warped random curves. *J. Amer. Statist. Assoc.* **99**, 687-699, where more details can be found.

In the presence of warping one faces simultaneous variation in amplitude and time and this often leads to identifiability issues. When each subject follows its own time scale, time synchronization as a pre-processing step often improves subsequent analysis of functional data. It is also of interest in itself. In gene time course expression analysis, gene classification can be based on simple time-shift warping (Silverman 1995, Leng and Müller 2006).

Landmark identification and alignment (Gasser and Kneip 1995) has become a gold standard for warping, especially for growth curves where the landmarks are often easy to identify (see Fig. 1). Landmarks have proved useful for the analysis of longitudinal growth curves early on (Gasser et al. 1984) due to the prominence of the growth spurts. In landmark warping one maps all landmark locations, often defined as peaks or troughs in first or second derivatives, to the average values across the sample

and interpolates the times in between for each individual (for example with an interpolating spline). Landmark methods however do not work in situations where the individual curves are variable to the extent that they do not share common shapes. Procrustes and area-under-the-curve registration are not subject to such shape constraints. Alternative robust warping methods have been developed lately (Gervini and Gasser 2005). Much work remains to be done in this area.

4.2 Methods for Time-Synchronization

Simple warping transformations include time-shift warping (Silverman 1995; Leng and Müller 2006) where one assumes in the simplest case for the i-th trajectory that $Y_i(t) = Y_0(t - \tau_i)$, τ_i denoting a (random) time shift for the i-th subject, and Y_0 a synchronized trajectory. Another simple variant that is often useful, especially when the sampled functions have varying domains, is scale warping. Here one models $Y_i(t) = Y_0(t/\sigma_i)$ for scale factors $\sigma_i > 0$. Both schemes can be combined, leading to shape-invariant modeling (Lindstrom 1995, Wang, Ke and Brown 2003).

A useful framework is to view warping as a time synchronization step, formalized as follows. Time for each subject is mapped from a standard or synchronized time $t \in [0, 1]$ to the individual or warped time $\tilde{X}(t)$, where this mapping must be strictly monotone and invertible, and in most approaches is considered to be a random function. Ideally, a warping method will satisfy the boundary conditions $\tilde{X}(0) = 0, \tilde{X}(1) = T$. The sample of observed trajectories can then be viewed as being generated by a latent bivariate stochastic process in "synchronized time space" $\mathcal{S}$ (Liu and Müller, 2004) $\{(\tilde{X}(t), \tilde{Y}(t)), t \in [0, 1]\} \subset L^2([0, 1]) \times L^2([0, 1])$. The observed sample then corresponds to $\{(\tilde{Y}(\tilde{X}^{-1}(x)), x \in [0, T]\} \subset L^2([0, T])$, and the associated warping mapping is

$$\psi : \{(\tilde{X}(t), \tilde{Y}(t)),\ t \in [0, 1]\} \mapsto \{(x, Y(x)),\ x \in [0, T]\},$$

defined by $Y(x) = \tilde{Y}(\tilde{X}^{-1}(x))$.

The identifiability problem corresponds to the fact that this mapping does not have a unique inverse. This is where the various warping methods such as Procrustes method or landmark warping come in, providing a concrete synchronization algorithm. A very simple but often effective warping method that is featured in Fig. 1 and 2 is area-under-the-curve warping. This works for samples of non-negative random trajectories. Here one assumes that synchronized time corresponds to the relative area under each individual trajectory. The total area is normalized to 1, and if the fraction of area under the curve is the same for two different observed times, these are considered to correspond to the same point in individual development and are mapped to the same synchronized time. Formally, to obtain the inverse warping process $\tilde{X}^{-1}$, which corresponds to the time-synchronizing mapping, as a function of each

observed trajectory processes Y, one simply determines the fractions of the area under the observed curves Y and defines this to be the synchronized time,

$$\varphi(Y)(x) = \tilde{X}^{-1}(x) = \frac{\int_0^x |Y(s)|\, ds}{\int_0^T |Y(s)|\, ds}.$$

Applying this time-synchronizing mapping is referred to area-under-the curve warping.

Considering the latent bivariate processes $\{(\tilde{X}(t), \tilde{Y}(t), t \in [0,1]\}$, as $\tilde{X}(\cdot)$ is constrained to be positive increasing, the space where the bivariate processes live is a convex space. This leads to a convex calculus. Given two observed processes Y_1, Y_2, and a fixed $0 \leq \pi \leq 1$, define a functional convex sum

$$\pi Y_1 \oplus (1-\pi) Y_2 = \psi\{\pi\psi^{-1}(Y_1) + (1-\pi)\psi^{-1}(Y_2)\},$$

where ψ^{-1} is the inverse mapping $\psi^{-1}(Y) = \{\{\varphi^{-1}(Y)\}(t), Y(\{\varphi^{-1}(Y)\}(t)),\, t \in [0,1]\}$. The functional convex sum can be easily extended to the case of K functions, $K > 2$,

$$\bigoplus_{j=1}^{K} \pi_j Y_j = \psi\big(\sum_{j=1}^{K} \pi_j X_j, \sum_{j=1}^{K} \pi_j Y_j\big).$$

This then leads to the warped average function (functional convex average, shown in Fig. 1 and 2)

$$\bar{Y}_{\oplus} = \bigoplus_{j=1}^{n} \frac{1}{n} Y_j.$$

Similarly, a convex path connecting observed random trajectories is $\{\pi Y_1 \oplus (1-\pi) Y_2, \quad \pi \in [0,1]\}$. Further results on this general warping framework and area-under-the-curve warping can be found in Liu and Müller (2003, 2004).

5 Functional Principal Component Analysis

5.1 Square Integrable Stochastic Processes

Functional Principal Component Analysis (FPCA) has emerged as a major tool for dimension reduction within FDA. One goals is to summarize the infinite-dimensional random trajectories through a finite number of functional principal component (FPC) scores. This method does not require distributional assumptions and is solely based on first and second order moments. It also provides eigenfunction estimates which are known as "modes of variation". These modes often have a direct biological interpretation and are of interest in their own right (Kirkpatrick and Heckman 1989). They offer a visual tool to assess the main directions in which the functional data vary. An important application is a representation of individual trajectories through an empirical Karhunen-Loève representation. It is always a good idea to check and adjust for warping before carrying out an FPCA.

For square integrable random trajectories $Y(t)$, we define mean and covariance functions

$$\mu(t) = E(Y(t)) \tag{4}$$

$$G(s,t) = \operatorname{cov}\{Y(s), Y(t)\}, \quad s,t \in \mathcal{T} \tag{5}$$

and the auto-covariance operator

$$(Af)(t) = \int_{\mathcal{T}} f(s)G(s,t)\,ds.$$

This is a linear Hilbert-Schmidt operator in the function space of square integrable functions $L^2(\mathcal{T})$ with Hilbert-Schmidt kernel G (Conway 1985). Under minimal assumptions this operator has orthonormal eigenfunctions ϕ_k, $k = 1, 2, \ldots$ with associated ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$, i.e., satisfying

$$A\,\phi_k = \lambda_k\,\phi_k.$$

The eigenfunctions of the auto-covariance operator turn out to be very useful in FDA for dimension reduction, due to the Karhunen-Loève expansion. This expansion holds under minimal assumptions (see Ash and Gardner 1975) and converges in the L^2 sense and also pointwise. It provides an important representation of individual trajectories Y,

$$Y(t) = \mu(t) + \sum_{k=1}^{\infty} A_k\,\phi_k(t), \tag{6}$$

where the A_k are uncorrelated random variables, known as the functional principal component scores (FPC scores). They satisfy $E(A_k) = 0, \quad \operatorname{var}(A_k) = \lambda_k$ and have the explicit representation

$$A_k = \int_{\mathcal{T}} (Y(t) - \mu(t))\phi_k(t)\,dt. \tag{7}$$

The situation is analogous to the representation of random vectors in multivariate analysis by principal components, replacing the scalar product in the vector space $\mathbf{R}^d$, given by $\langle x, y\rangle = \sum_{k=1}^{d} x_k y_k$ by $\langle x, y\rangle = \int x(t)y(t)\,dt$, and replacing matrices by linear operators; however since we deal with infinite sums, issues of convergence arise.

5.2 From Karhunen-Loève Representation to Functional Principal Components

One of the main attractions of FPCA is the equivalence (in distribution) of $Y - \mu$ (centered process) and the (uncorrelated) FPC scores, $Y - \mu \equiv \{A_1, A_2, \ldots\}$. This is a consequence of the Karhunen-Loève representation. In applications, the sequence of FPC scores is truncated at a suitable index (if possible, chosen data-adaptively). This truncation constitutes the needed regularization step, mapping

the infinite trajectories to a finite number of FPC scores. Along the way, one also needs to estimate the (smooth) mean functions and the relevant eigenfunctions. This can be done by smoothing methods as demonstrated below.

Alternative representations of functional data by expansions in fixed basis functions have also been considered. These include Fourier and wavelet bases (Morris and Carroll 2006). These representations have the advantage that neither mean nor eigenfunctions need to be determined from the data. Wavelets are particularly suited for data with somewhat non-smooth trajectories such as functions containing small jumps and sharp edges. They are less well suited to reproduce really smooth trajectories. The disadvantage of fixed basis functions is that they may not be very parsimonious and many more basis functions are needed to represent a given sample of trajectories. In addition, the estimated coefficients are not uncorrelated (which means they carry less information and are less convenient for subsequent applications such as regression).

A preliminary exploration of functional principal components for longitudinal data is due to C.R. Rao (1958). Other key references are Castro, Lawton and Sylvestre (1987), who introduced the notion that eigenfunctions are functional "modes of variation", Rice and Silverman (1991), who emphasized the need for smoothing for which they used B-splines, and Ramsay and Silverman (2005), who start with a pre-smoothing step to first generate a sample of smooth trajectories, before proceeding with FPCA.

If complete trajectories are observed, or data observed on a grid are pre-smoothed and then considered as completely observed, one typically creates an equidistant grid $\{t_1, t_2, \ldots, t_N\}$ of N design points on the domain $\mathcal{T}$ (where N is the same as the number of sampled trajectories, which corresponds to the number of subjects) and then one treats the data as N-vectors, one for each of the N subjects. One then performs a multivariate principal component analysis for these N-vectors, i.e., one obtains mean vector, eigenvectors and principal component scores, without any smoothing (compare Cardot, Ferraty and Sarda 1999). Theoretical analysis then focuses on asymptotics as $N \to \infty$. If data however are not densely or irregularly sampled, or are contaminated with noise, this approach does not work and smoothing is necessary. As noise-contaminated measurements are rather the norm than the exception, the case of completely observed trajectories is mainly of theoretical interest.

5.3 The Case of Longitudinal Data

For sparsely sampled longitudinal data, pre-smoothing to create completely observed trajectories is a less attractive option as it introduces bias and artificial correlations into longitudinal data. This

is because scatterplot smoothing requires relatively dense and not too irregular designs. If there are "gaps" in the predictors, bandwidths must be increased which in turn leads to increased bias. Irregular and sparse data, as typically encountered in longitudinal studies, were first considered by James, Hastie and Sugar (2001), who used B-splines. The B-spline approach with random coefficients, pioneered by Shi, Weiss and Taylor (1996) and Rice and Wu (2000), can also be easily adapted to the sparse and irregular case.

The FPCA approach and Karhunen-Loève representation cannot be directly adopted to longitudinal data, which from now on we assume to consist of sparse, irregular and noisy measurements of the longitudinal trajectories. According to (7), the FPC scores which are the random effects in the representation would be estimated by approximating the integral by a Riemann sum. This works nicely for the case of fully observed trajectories but does not work for longitudinal data, due to large discretization errors. If the data are contaminated by noise, the approximation by sums does not work consistently, even if the measurements are dense. The case of noisy measurements in FDA was first emphasized in the work of Staniswalis and Lee (1998).

We model noisy longitudinal data as follows: Let Y_{ij} be measurements of trajectories $Y_i(\cdot)$ made at sparse and irregularly spaced time points t_{ij}, $1 \leq i \leq n, 1 \leq j \leq n_i$. Then

$$Y_{ij} = Y_i(t_{ij}) + \epsilon_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} A_{ik}\phi_k(t_{ij}) + \epsilon_{ij}.$$

Here the ϵ_{ij} are i.i.d. measurement errors with moments $E\epsilon_{ij} = 0$, $E\epsilon_{ij}^2 = \sigma^2$, and the ϵ_{ij} are considered to be independent of the FPC scores A_{ik}, denoting the score for the $i-$th subject and $k-$th eigenfunction.

An example for sparse and irregular data for which this model may apply are longitudinal CD4 counts of AIDS patients (Fig. 3, left panel).

5.4 Principal Analysis by Conditional Expectation

Her we describe the PACE (Principal Analysis via Conditional Expectation) method to carry out FPCA for longitudinal (used synonymously here for sparse and irregularly sampled) data (Yao et al. 2005a). The basis for this method is the PART – Principal Analysis of Random Trajectories algorithm for obtaining the empirical Karhunen-Lovè representation of smooth functional data, where measurements are contaminated with additional measurement error. This algorithm works irrespective of whether the measurements have been sampled or a dense and regular grid or on a sparse and irregular grid. Alternative algorithms that use pre-smoothing are available (see, e.g., Ramsay and Silverman 2005 and the associated web site).

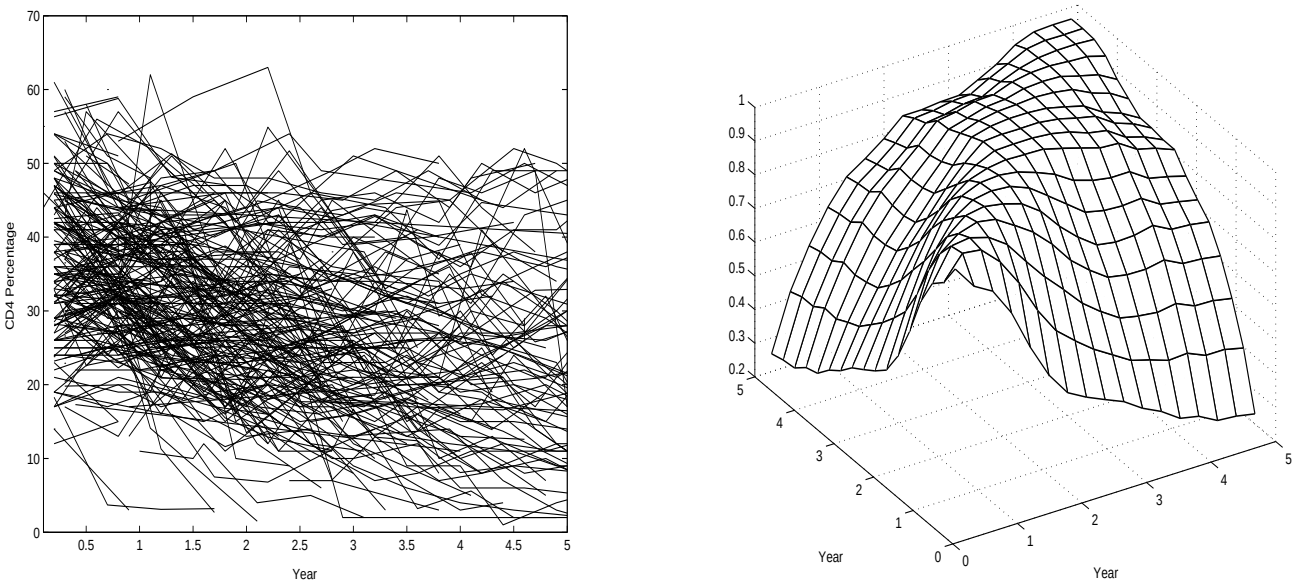

Figure 3: Longitudinal CD4 Data for $N = 238$ male subjects, here shown for a subsample of 25 subjects. Left panel: Plot of the data, connecting repeated measurements by straight lines. Right panel: Smooth estimate of the covariance surface, where the diagonal has been removed. Figure reproduced from Yao, F., Müller, H.G., Wang, J.L. (2005). Functional data analysis for sparse longitudinal data. *J. American Statistical Association* **100**, 577-590.

The PART algorithm consists of the following steps: In a first step, one pools all available measurements $(t_{ij}, Y_{ij}), \quad i = 1, \ldots, N$, $j = 1, \ldots, n_i$, into one scatterplot and uses a one-dimensional smoother to obtain the estimate $\hat{\mu}(t)$ of the overall mean function $\mu(t)$. A technical requirement here is that the pooled locations t_{ij} over all subjects are dense on the domain or at least can be reasonably considered to become dense asymptotically. This will lead to consistency for this estimation step. Next, one forms all pairwise products

$$(Y_{ij} - \hat{\mu}(t_{ij}))(Y_{il} - \hat{\mu}(t_{il})), \quad j \neq l,$$

which will be the responses for predictors (t_{ij}, t_{il}). Both these responses and the predictor are entered into a 2-dimensional scatterplot smoother. The output is the estimated covariance surface. The diagonal elements (for which $j = l$) are omitted from the input into the 2-dimensional smoother, since they are contaminated by the measurement errors. The measurement error variance in fact can be estimated from these diagonal elements, either under the assumption that it is a fixed constant, or under the assumption that it varies over time. In the latter case, one obtains the variance function of the errors by smoothing along the direction of the diagonal.

While in our implementation we use the smoothers described in subsection 1.2, any alternative smoothing method can be used as well. One potential problem is that while the estimated covariance matrix is easily seen to be symmetric (as the responses that are entered are symmetric in t_{ij}, t_{il}), it is not necessarily positive definite. This problem can be solved by projecting on positive definite surfaces, truncating negative eigenvalues (for details, see Yao et al. 2003). From the estimated covariance surface, one obtains eigenfunctions and eigenvalues numerically after discretizing. The bandwidths for the smoothing steps can be obtained by cross-validation or similar procedures.

Once mean function and eigenfunctions have been obtained, an important step in completing the empirical Karhunen-Loève representation, and thus the functional dimension reduction, is the estimation of the FPC scores. Following representation (7) of the scores, one could plug estimates for the unknown quantities into the integral which could be approximated by a Riemann sum to obtain estimated FPC scores. This is however only reasonable for sufficiently dense designs and for noise-free observations. If the observations are noisy or sparse, the Riemann sums will not provide reasonable approximations to the integral.

This is where the PACE approach to predict individual FPC scores comes in. We need Gaussian

assumptions, i.e., A_{ik}, ϵ_{ij} are jointly normal. Define

$$\begin{aligned}\widetilde{Y}_i &= (Y_{i1}, \cdots, Y_{in_i})^T, \\ \mu_i &= (\mu(t_{i1}), \cdots, \mu(t_{iN_i}))^T, \\ \phi_{ik} &= (\phi_k(t_{i1}), \cdots, \phi_k(t_{iN_i}))^T.\end{aligned}$$

The best predictors for the random effects are obtained via the conditional expectation

$$\begin{aligned}E[A_{ik}|\widetilde{Y}_i] &= E(A_{ik}) + \mathrm{cov}(A_{ik}, \widetilde{Y}_i)\mathrm{cov}(\widetilde{Y}_i, \widetilde{Y}_i)^{-1}(\widetilde{Y}_i - \mu_i) \\ &= \lambda_k \phi_{ik}^T \Sigma_{Y_i}^{-1}(\widetilde{Y}_i - \mu_i),\end{aligned}$$

where

$$(\Sigma_{Y_i})_{j,l} = \mathrm{cov}(Y_i(t_{ij}), Y_i(t_{il})) + \sigma^2 \delta_{jl}, \quad \delta_{jl} = 1 \text{ if } j = l, \text{ and } 0 \text{ if } j \neq l.$$

Plugging in the estimates discussed above then leads to estimated predicted FPC scores

$$\hat{E}[A_{ik}|\widetilde{Y}_i] = \hat{\lambda}_k \hat{\phi}_{ik}^T \hat{\Sigma}_{Y_i}^{-1}(\widetilde{Y}_i - \hat{\mu}_i). \tag{8}$$

5.5 Predicting Individual Trajectories

Once the number of included random coefficients K has been determined, we can use the predicted FPC scores (8) to obtain predicted individual trajectories

$$\hat{Y}(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{E}(A_k|\tilde{Y}_i)\, \hat{\phi}_k(t). \tag{9}$$

An important issue is the choice of K, the number of included components. This corresponds to the number of FPC scores and accordingly, the number of random effects in the model. For this choice, one can use the scree plot. This plots the fraction of variance unexplained by the first K components as a function of K,

$$S(K) = 1 - \sum_{k=1}^{K} \hat{\lambda}_k / \sum_{k=1}^{\infty} \hat{\lambda}_k.$$

One looks for a "knee" in this graph, i.e., a value of K at which the rate of decline slows substantially, as K increases further.

A second promising approach are AIC-type criteria. As no likelihood exists a priori, one can devise various types of pseudo-likelihood and then construct a pseudo-AIC value. For example, a pseudo-Gaussian log-likelihood is

$$\widehat{L}_1 = \sum_{i=1}^{N} \{-\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log(\det\widehat{\Sigma}_{\tilde{Y}_i}) - \frac{1}{2}(\tilde{Y}_i - \hat{\mu}_i)^T \widehat{\Sigma}_{\tilde{Y}_i}^{-1}(\tilde{Y}_i - \hat{\mu}_i)\}, \tag{10}$$

while a conditional version of the likelihood, conditioning on predicted FPC scores, would be

$$\widehat{L} = \sum_{i=1}^{N} \{ -\frac{n_i}{2} \log(2\pi) - \frac{n_i}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} (\widetilde{Y}_i - \hat{\mu}_i - \sum_{k=1}^{K} \hat{A}_{ik} \hat{\phi}_{ik})^T (\widetilde{Y}_i - \hat{\mu}_i - \sum_{k=1}^{K} \hat{A}_{ik} \hat{\phi}_{ik}) \}. \tag{11}$$

In either version, the pseudo-AIC value is then $\text{AIC} = -\widehat{L} + K$.

A characteristic of the PACE method is that it borrows strength from the entire sample to predict individual trajectories, in contrast to the more traditional nonparametric regression analysis of longitudinal data, where each curve would be fitted separately from the others by appropriate smoothing. We note that this traditional approach has been successful for regular designs as encountered in growth studies, but is less feasible for the more typical longitudinal data where the number of observations per curve is small. In theoretical analysis this is adequately reflected by assuming that the number of repeated measurements per subject is bounded, while the number of individuals will potentially be large. We note that once the FPC scores have been computed, they can be entered into further statistical analysis. Pairwise scatterplots of one FPC score against another, plotted for all subjects, can reveal patterns of interest. Pairs or vectors of FPC scores are useful for classification or clustering of samples of trajectories (Müller 2005; compare also James and Sugar 2003).

A number of asymptotic properties of functional principal components have been investigated. Most of the earlier results are due to the French school (Dauxois, Pousse and Romain 1982). Assuming more than one but at most finitely many observations are available per trajectory, and without assuming Gaussian assumptions, it was shown in Hall, Müller and Wang (2006) that the eigenfunction estimates achieve the usual nonparametric rates for estimating a smooth function, as sample size $N \to \infty$. For the case where entire trajectories are available without measurement error, the optimal rates are parametric (Hall & Hossini-Nasab 2006). The above estimates for covariance surface and mean function converge in sup-norm and so do the eigenfunctions, under longitudinal designs (Yao et al. 2005a). The predicted FPC scores converge to the actual FPC scores as the designs get denser (more and more measurements per subject, see Müller 2005). Under additional Gaussian assumptions, then the estimates of the predicted FPC scores converge to their targets, and pointwise/uniform confidence bands can be constructed for predicted trajectories (Yao et al 2005a).

5.6 Application to Longitudinal CD4 Data

As an illustration, this algorithm was applied to longitudinal CD4 counts for a sample of CD4 counts obtained for 283 male AIDS patients (Multicenter AIDS Cohort Study 1987). Potential issues with informative drop-out in this study are ignored in this analysis. The data fit the description of sparse

and irregular data and are shown in the left panel of Fig. 3, where the data for each individual are connected by straight lines. The numbers of observations per subject are between 2 and 14. We aim at describing the characteristic features of the underlying longitudinal trajectories. The estimated covariance function for these data is in the right panel of Fig. 3, where the diagonal has been omitted as described above. The overall mean function $\hat{\mu}(t)$ is depicted in the lower panels of Fig. 4. The conditional pseudo-likelihood (11) based AIC criterion yielded $K = 3$, i.e., three eigenfunctions are included.

Of interest is an assessment of the extremes in a sample. In the functional situation it is not so straightforward what these extremes are. One possibility is to identify those subjects whose trajectories are most aligned with an eigenfunction. This device and further exploration of samples of trajectories by means of the eigenfunctions has been studied by Jones and Rice (1992). The lower panels of Fig. 4 display these extreme trajectories. These trajectories and the eigenfunctions provide an idea about the modes of variation that are present. The first mode is a linear decline, exemplified by the subject in the left lower panel of Fig. 4; the second mode is a decline with a plateau in the middle, during the third year, after which a more rapid decline in CD4 counts resumes. The third and weakest mode corresponds to a leveling off towards the end, stabilizing at a low level, with a possible increase. One should not read too much into the increase at the right end of the third eigenfunction; this may simply be caused by boundary effects.

Finally, we are interested to model individual trajectories, which are obtained via the estimated FPC scores, see equation (9). The predicted trajectories for four subjects, including confidence bands, are shown in Fig. 5, including Gaussian-based confidence bands. Open questions that will be of interest for future research include the extension of FDA methods to repeated non-Gaussian (binomial, Poisson) data, the case of varying eigenfunctions in dependency on a covariate, and incorporating informative missingness and time-to-event information. A full exploration of practical features in the context of various longitudinal studies will also be of interest.

6 Functional Regression Models

6.1 Overview

For longitudinal data analysis, the trajectories observed for each subject can serve as both predictors and response in a regression model. The case where they are included among the predictors has been well explored in FDA, primarily for the case where the trajectories are fully observed without noise (Cardot et al. 2003, Cai and Hall 2006). We review here some of the available models. Linear functional models

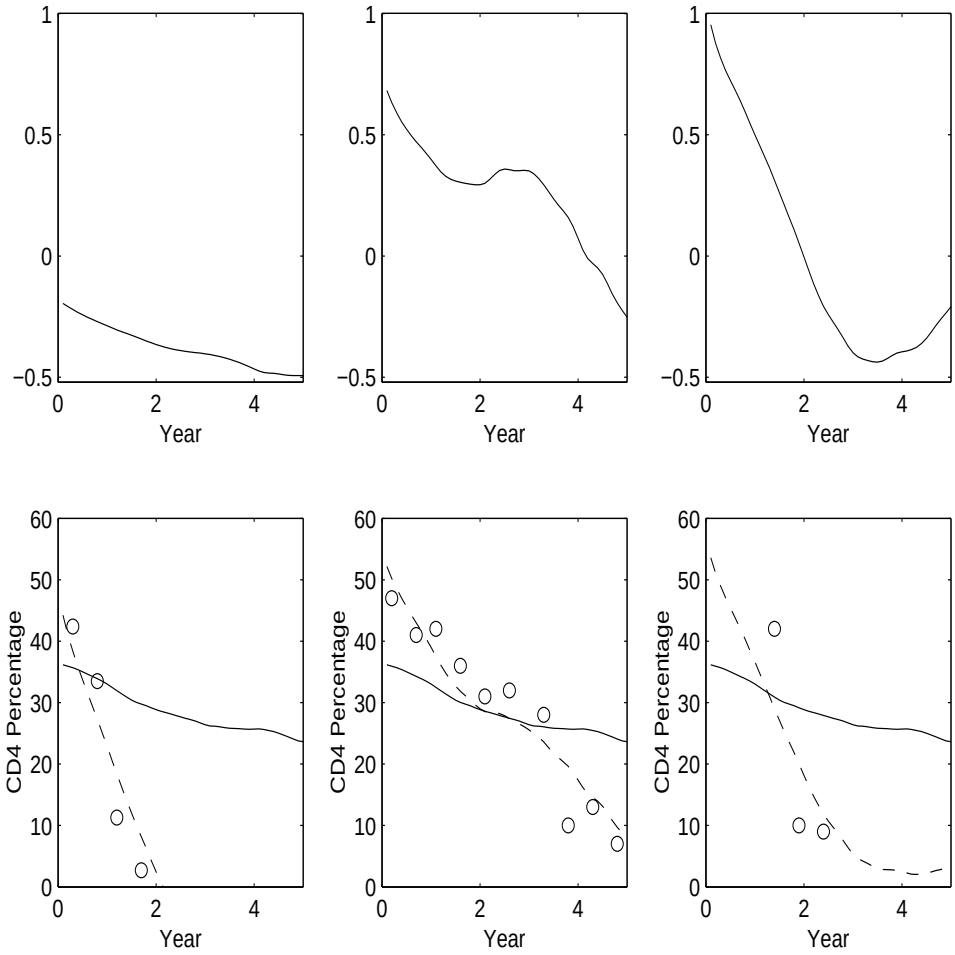

Figure 4: Eigenfunctions, mean function and extreme trajectories for longitudinal CD4 data. Upper three panels: First, second and third eigenfunction, from left to right. Lower three panels: Mean function for all trajectories (solid), and three individuals with data and fitted trajectories, most aligned in the directions of the three eigenfunctions. Figure reproduced from Yao, F., Müller, H.G., Wang, J.L. (2005). Functional data analysis for sparse longitudinal data. *J. American Statistical Association* **100**, 577-590.

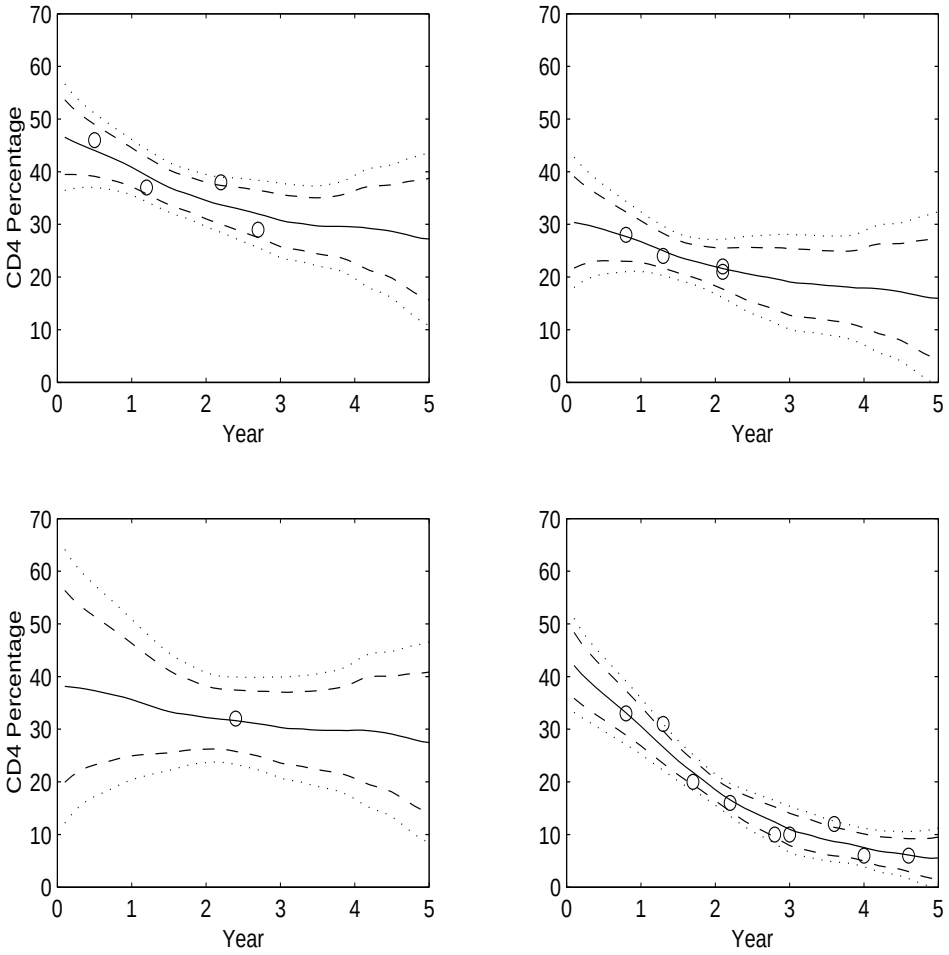

Figure 5: Predicted trajectories for four subjects of the longitudinal CD4 study. Each panel displays the observed data (circles), predicted trajectory (solid) and local (dashed) as well as uniform (dotted) 95% confidence bands, based on the Gaussian assumption. Figure reproduced from Yao, F., Müller, H.G., Wang, J.L. (2005). Functional data analysis for sparse longitudinal data. *J. American Statistical Association* **100**, 577-590.

may include a random trajectory in either predictors, responses, or both. We assume here the data are written as (X, Y), where Y stands for response and X for predictor, which could be scalar or functional. Means will be denoted by μ_X, μ_Y. In the functional case we denote eigenfunctions by ϕ_k for X and ψ_k for Y.

The linear model for a scalar response and a functional predictor is

$$E(Y|X) = \mu_Y + \int_{\mathcal{T}} (X(s) - \mu_X(s))\beta(s)\, ds,$$

where β is the regression parameter function. An extension to functional responses is the model

$$E(Y(t)|X) = \mu_Y(t) + \int_{\mathcal{T}} (X(s) - \mu_X(s))\beta(s,t)\, ds, \tag{12}$$

where now the regression parameter function has two arguments, i.e., is a surface. This model dates back to Ramsay and Dalzell (1991). It can be interpreted as an extension of the multivariate linear regression model $E(Y|X) = BX$ for a parameter matrix B to the functional case.

In such a multivariate linear regression model a common estimation scheme proceeds via the least squares normal equation: For $X \in \mathbf{R}^p, \quad Y \in \mathbf{R}^q$, the normal equation is $\mathrm{cov}(X, Y) = \mathrm{cov}(X)B$, where $\mathrm{cov}(X, Y)$ is the $p \times q$ matrix with elements $a_{jk} = \mathrm{cov}(X_j, Y_k)$. This equation can be solved for B if the $p \times p$ covariance matrix $\mathrm{cov}(X)$ is invertible. The situation is much less straightforward for the functional extension. We can define an analogous "Functional Normal Equation" (He, Müller and Wang 2000),

$$r_{XY} = R_{XX}\beta, \quad \text{for} \quad \beta \in L_2,$$

where $R_{XX} : L^2 \to L^2$ is the auto-covariance operator of X, defined by

$$(R_{XX}\beta)(s,t) = \int r_{XX}(s,w)\beta(w,t)dw,$$

where

$$r_{XX}(s,t) = \mathrm{cov}\left[X(s), X(t)\right], \quad r_{XY}(s,t) = \mathrm{cov}\left[X(s), Y(t)\right].$$

As R_{XX} is a compact operator in L^2, it is in principle not invertible. Thus we face an inverse problem which requires regularization (compare He, Müller and Wang 2003).

A model that is useful in classifying longitudinal time courses is the generalized functional linear model (James 2002; Müller and Stadtmüller 2005; Müller 2005). Here the predictors are functional, the responses are generalized variables such as binary outcomes which may stand for class membership or Poisson counts. With an appropriate link function g, this model can be written as

$$E(Y|X) = g\left(\mu + \int_{\mathcal{T}} X(s)\beta(s)\, ds\right), \tag{13}$$

coupled with a variance function $\mathrm{var}(Y|X) = V(E(Y|X))$. This model is an extension of the common Generalized Linear Model. It can be implemented with both known or unknown link/variance function (see Müller and Stadtmüller 2005).

The class of "functional response models" (Faraway 1997, Chiou, Müller and Wang 2003, 2004) is of interest in functional dose-response models and similar applications. In this model the predictor is usually a vector Z, while the response is functional,

$$E\{Y(t)|Z = z\} = \mu(t) + \sum_{k=1}^{K} \alpha_k(\gamma_k^T z)\psi_k(t).$$

Here the γ_k are single indices (i.e., vectors which project the covariates Z onto one dimension, and the α_k are link functions to the random effects. Sometimes simpler structured models such as a "multiplicative effects model" are useful,

$$\mu(t, z) = \mu_0(t)\theta(z),\ E\{Y(t)\} = \mu_0(t),\ E(\theta(Z)) = 1,$$

for a function $\theta(\cdot)$ (see Chiou et al 2003).

Further classes of models of interest are those with varying supports. In the regression models above the entire predictor function is assumed to contribute to a response. In many applications this might not be realistic. Examples for this were given in Malfait and Ramsay and Müller and Zhang (2005). In the latter paper, the response is remaining lifetime, to be predicted from a longitudinal covariate which is observed to current time. As current time progresses, the functional regression model needs to be updated. This leads to time-varying domains and accordingly to time-varying coefficient functional regression models. In the extreme case, the usual varying coefficient model

$$E(Y(t)|X) = \mu_Y(t) + \beta(t)X(t)$$

(under the assumption of one predictor process) emerges as a special case; here $\beta(\cdot)$ is the varying coefficient function (Fan and Zhang 1998, Wu and Yu 2002).

These models can be extended to the longitudinal (sparse and irregular) case, following the above PACE approach whenever the model can be written in terms of FPC scores. In the following we show this for the functional regression model (12).

6.2 Functional Regression for Longitudinal Data

Extending the functional linear regression model (12) introduced above for FPCA to the case of sparse and irregular data, we assume that available measurements for predictor and response curves are given

as follows, with their Karhunen-Loève representations included,

$$U_{il} = X_i(s_{il}) + e_{il} = \mu_X(s_{il}) + \sum_{m=1}^{\infty} A_{im}\phi_m(s_{il}) + e_{il},$$
$$V_{ij} = Y_i(t_{ij}) + \epsilon_{ij} = \mu_Y(t_{ij}) + \sum_{k=1}^{\infty} B_{ik}\psi_k(t_{ij}) + \epsilon_{ij},$$

where the times where measurements s_{ij}, recorded for predictor processes X, resp. t_{ij}, recorded for response processes Y, can differ between X and Y, but are both assumed to be sparse. The random effects (FPCA scores) are denoted here by A_{im} for predictor processes and by B_{ik} for response processes.

Applying FPCA, by using the orthonormality properties of the eigenfunctions, one finds that the regression parameter function β in (12) can be represented by

$$\beta(s,t) = \sum_{k,m=1}^{\infty} \frac{E[A_m B_k]}{E[A_m^2]}\phi_m(s)\psi_k(t). \tag{14}$$

This reduces the problem to estimate β to the problem to obtain an estimate of $E[A_m B_k]$, for which we consider

$$\widehat{E}[A_m B_k] = \int_0^{\mathcal{T}} \int_0^{\mathcal{S}} \hat{\phi}_m(s)\widehat{\Gamma}_{XY}(s,t)\hat{\psi}_k(t)\, dsdt, \tag{15}$$

where $\widehat{\Gamma}_{XY}(s,t)$ is a local linear smoother for the cross-covariance function $\Gamma_{XY}(s,t) = \mathrm{cov}(X(s), Y(t))$ (Yao et al. 2005b).

Once the regression parameter surface β has been obtained, one then may aim at predicting individual response trajectories, from the available observations of the corresponding predictor process, i.e., to predict Y^* from the observations $U^* = (U_1^*, \cdots, U_{L^*}^*)^T$ available for $X^*(\cdot)$, where $*$ means these are the data for one individual. Under Gaussian assumptions, the best predictor is given by

$$\begin{aligned} E[Y^*(t)|X^*(\cdot)] &= \mu_Y(t) + \int_0^{\mathcal{S}} \beta(s,t)(X^*(s) - \mu_X(s)\, ds \\ &= \mu_Y(t) + \sum_{k,m=1}^{\infty} \frac{E[A_m B_k]}{E[A_m^2]} A_m^* \psi_k(t). \end{aligned}$$

An estimate for this predictor is simply obtained by plugging in estimates for the unknown quantities. Choosing K and M for the number of included components to represent $X-$ and $Y-$ processes, we arrive at

$$\widehat{Y}_{KM}^*(t) = \hat{\mu}_Y(t) + \sum_{m=1}^{M}\sum_{k=1}^{K} \frac{\widehat{E}[A_m B_k]}{\widehat{E}[A_m^2]} \widehat{E}[A_m^*|U^*]\hat{\psi}_k(t), \tag{16}$$

where $\widehat{E}[A_m^*|U^*]$ is estimated by the PACE method as described in subsection 4.3, given observations $U^* = (U_1^*, \cdots, U_{L^*}^*)^T$ of $X^*(\cdot)$.

Theory developed in Yao et al (2005b) includes consistency of the regression parameter surface estimates, as well as some basic inference, and also construction of pointwise and uniform confidence bands for predicted trajectories, under Gaussian assumptions. This paper also contains extensions of the usual coefficient of determination $R^2 = \text{var}(E[Y|X])/\text{var}(Y)$ to the functional case. Applying orthonormality properties of the eigenfunctions, one such possible extension can be represented as

$$R^2 = \frac{\int_{\mathcal{T}} \text{var}(E[Y(t)|X])dt}{\int_{\mathcal{T}} \text{var}(Y(t))dt} = \frac{\sum_{k,m=1}^{\infty} E(A_m B_k)^2 / E(A_m)^2}{\sum_{k=1}^{\infty} E(B_k)^2}. \tag{17}$$

The quantities $E(A_m)^2$, $E(B_k)^2$ correspond to the eigenvalues of the $X-$ and $Y-$ processes, and $E(A_m B_k)$ can be estimated as in (15), which then leads to estimates of this version of functional R^2.

6.3 Illustration with Data from the Baltimore Longitudinal Study of Aging

Longitudinal measurements of Body mass index (BMI) and systolic blood pressure (SBP) were obtained for 812 participants of the Baltimore Longitudinal Study on Aging (BLSA), as reported in (Pearson et al. 1997). The measurements fit the description of irregular and sparse. We provide a brief summary of the functional regression analysis conducted in Yao et al (2005b). The data and mean function estimates for all subjects can be found in Fig. 6. From this figure one can see the irregular nature of the timings as well as their sparseness. The relationship between the trajectories in left and right panels is difficult to discern.

Running the functional regression machinery, we obtain the estimate of the regression surface function $\hat{\beta}(\cdot, \cdot)$ for these data as depicted in Fig. 7. This function illustrates the influence of predictor functions on response trajectories. The time axis of predictor trajectories is labeled s, running towards the right, while the time axis of response trajectories is labeled t, running towards the left. In this functional regression model, the the entire predictor trajectory influences the entire response curve. We can interpret the features of this regression parameter surface as follows: At early ages, around 60, SBP is related to an overall average of BMI. At late ages, around 80, SBP is positively correlated with what is best characterized as rate of increase in BMI. A continuous transition between these regimes occurs in between.

Finally, predicted trajectories of systolic blood pressure for four randomly selected participants are displayed in Fig. 8. The predictors are the measurements of body mass index which are not shown in the graphs. A curious stricture occurs in the confidence bands around age 75. This is an area where apparently the variation of the response trajectories has a minimum.

The methods described above can be extended to the case of more than one predictor process, where one can use the FPC scores derived from the different predictor processes as predictors.

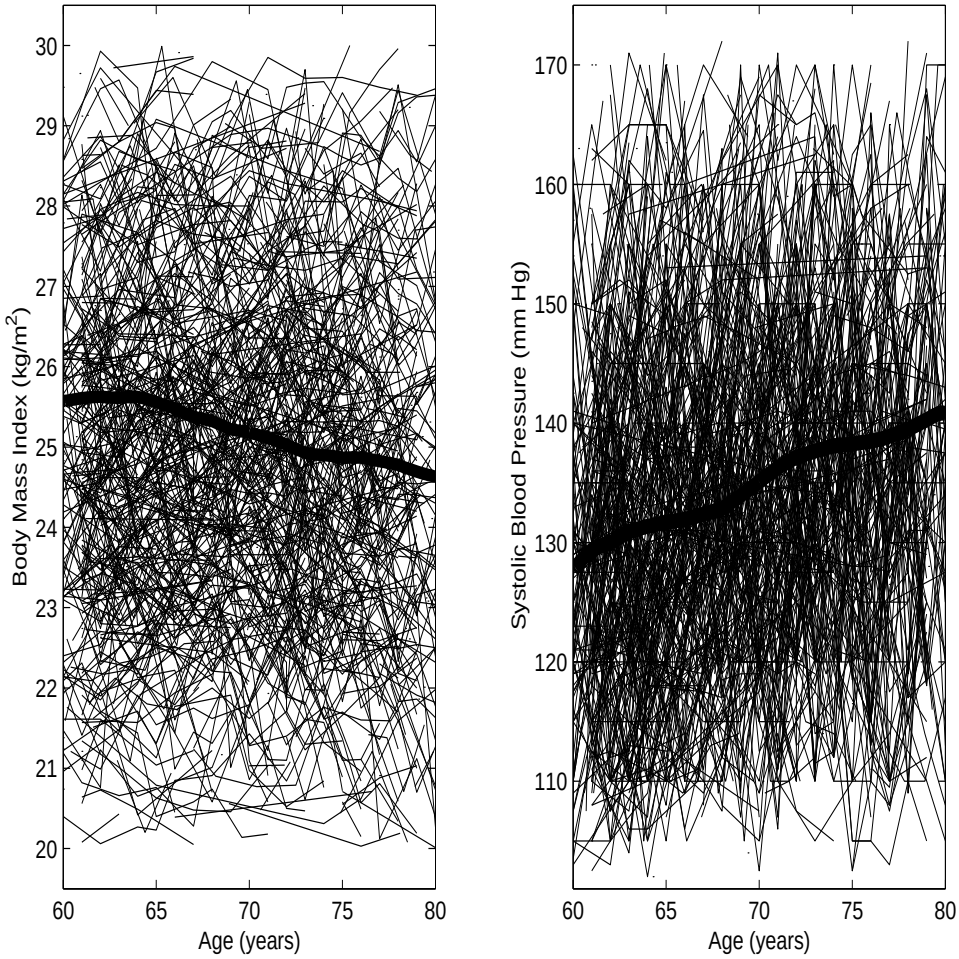

Figure 6: Longitudinal measurements of body mass index (left panel) and systolic blood pressure (right panel) for 812 participants in the Baltimore Longitudinal Study of Aging (BLSA). Thick solid curves are estimated mean functions. Reproduced from the article Yao, F., Müller, H.G., Wang, J.L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics* **33**, 2873-2903.

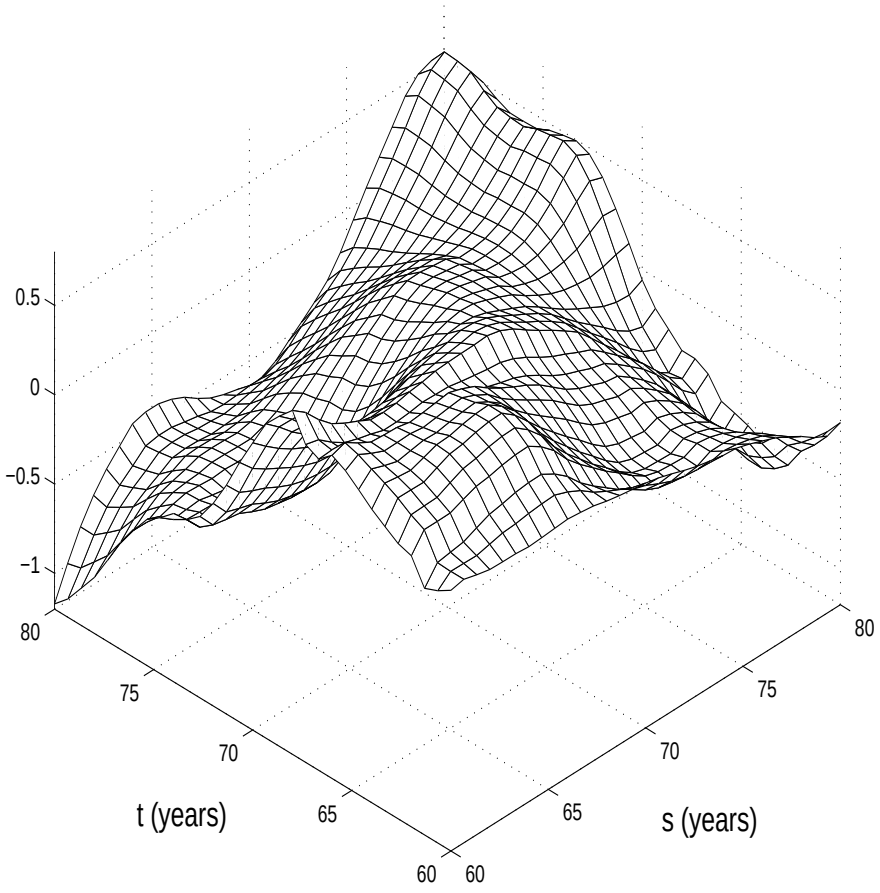

Figure 7: Estimated regression parameter surface β (14) for BLSA data. Reproduced from the article Yao, F., Müller, H.G., Wang, J.L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics* **33**, 2873-2903.

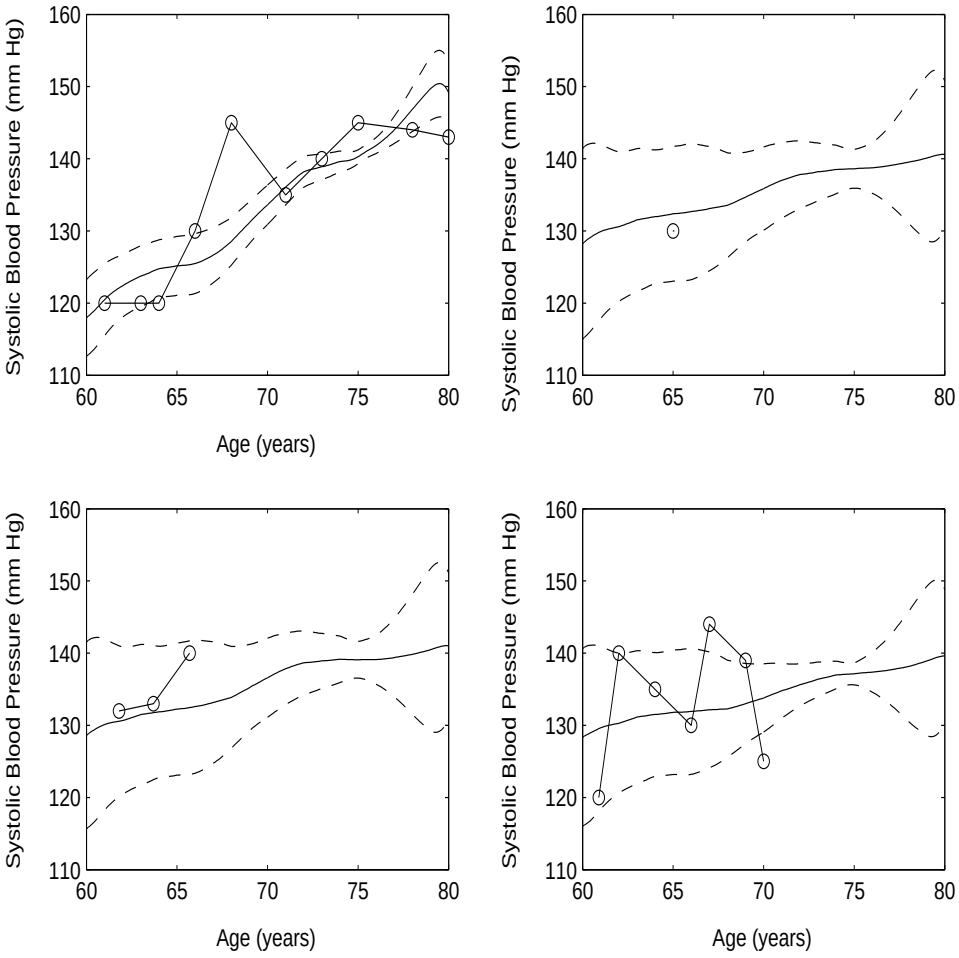

Figure 8: Predicted response trajectories for four participants of the BLSA study. Shown are predicted trajectories (solid curves), data for the response trajectories (circles, connected by straight lines) and pointwise 95% confidence intervals. Note that the data shown are not used for the prediction which is entirely based on measurements relating to predictor trajectories. Reproduced from the article Yao, F., Müller, H.G., Wang, J.L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics* **33**, 2873-2903.

7 Outlook

Besides the functional data analysis methodology described above, several approaches to functional ANOVA have applications for longitudinal data. These include ANOVA decompositions using smoothing spline smoothing, proposed in Brumback and Rice (1998) and applications of P-splines (Bugli and Lambert 2006). Other non- or semiparametric models of interest for longitudinal studies are varying coefficient models where $Y(t)$ is related to a series of predictors $X_1(t), \ldots, X_p(t)$. Typically one conducts a linear regression at each time point in a grid of time points and then one smooths the resulting regression coefficients. Furthermore, shape-invariant modeling is a promising functional method.

For functional inference, bootstrap based on the data of one individual as a resampling unit is being used but the theoretical foundations have not yet been developed. Asymptotic inference is available under Gaussian assumptions, but is not available on a wider scale; see Fan and Lin (1998). Functional approaches provide a flexible alternative to common parametric models for analyzing longitudinal data and are computationally easy to implement.

At this time, a number of key techniques are in place, notably smoothing and differentiation of noisy data, warping, functional principal components and penalized regularization. The unique combination of techniques from functional analysis, stochastic processes, nonparametrics and multivariate analysis and the many open questions make this a rewarding research area.

8 References

Ash, R.B. & Gardner, M.F. (1975). *Topics in Stochastic Processes.* New York: Academic Press.

Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Oxford Science Publications, Oxford.

Brumback B. and 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.

Bugli, C. and Lambert, P. (2006). Functional ANOVA with random functional effects: an application to event-related potentials modelling for electroencephalograms analysis. *Statist. Med.* to appear.

Cai, T. & Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics*

Capra, W.B. & Müller, H.G. (1997). An accelerated-time model for response curves. *J. Amer. Statist. Assoc.*, **92**, 72-83.

Cardot, H., Ferraty, F. & Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters* **45**, 11-22.

Cardot, H., Ferraty, F., Mas, A. & Sarda, P. (2003). Testing hypotheses in the functional linear model. *Scand. J. Statist.* **30**, 241-255.

Castro, P.E., Lawton, W.H. & Sylvestre, E.A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28**, 329-337.

Chiou, J.M., Müller, H.G. (2005). Estimated estimating equations: semiparametric inference for clustered/longitudinal data. *J. Royal Statistical Society B* **67**, 531-553.

Chiou, J. M., Müller, H. G. & Wang, J. L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J. Royal Statist. Assoc. Series B* **65**, 405-423.

Chiou, J.M., Müller, H.G. & Wang, J.L. (2004). Functional response models. *Statistica Sinica* **14**, 675-693.

Chiou, J.M., Müller, H.G., Wang, J.L., Carey, J.R. (2003). A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies. *Statistica Sinica* **13**, 1119-1133.

Conway, J. B. (1985). *A Course in Functional Analysis.* New York: Springer-Verlag.

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

Dubin, J. & Müller, H.G. (2005). Dynamical correlation for multivariate longitudinal data. *J. American Statistical Association* **100**, 872-881.

Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing.* Marcel Dekker.

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall, London.

Fan, J. & Lin S. K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93**, 1007-1021.

Fan, J. & Zhang, J. T. (1998). Functional linear models for longitudinal data. *J. Royal Statist. Assoc. Series B* **39**, 254-261.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics* **39**, 254-262.

Ferraty, F. & Vieu, P. (2006). Nonparametric Functional Data Analysis. Springer, New York.

Gasser, T. & Kneip, A. (1995). Searching for structure in curve samples. *J. Amer. Statist. Assoc.*, **90**, 1179-1188.

Gasser, T. & Müller, H.G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian J. Statistics* **11**, 171-184.

Gasser, T., Müller, H.G., Köhler, W., Molinari, L. & Prader, A. (1984). Nonparametric Regression Analysis of Growth Curves. *Ann. Statist.* **12**, 210-229.

Gasser, T., Müller, H.G., Köhler, W., Prader, A., Largo, R. & Molinari, L. (1985). An analysis of the mid-growth spurt and of the adolescent growth spurt based on acceleration. *Ann. Human Biology* **12**, 129-148.

Gasser, T., Müller, H.G., Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Royal Statistical Society B* **47**, 238-252.

Gervini, D. & Gasser, T. (2004). Self-modeling warping functions. *Journal of the Royal Statistical Society B* **66**, 959-971.

Gervini, D. & Gasser, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92**, 801-820.

Hall, P. & Hosseini-Nasab, M. (2006). On properties of functional principal component analysis. *J. Royal Statistical Society B* **68**, 109-126.

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

He, G., Müller, H.G. & Wang, J.L. (2000). Extending correlation and regression from multivariate to functional data. *Asymptotics in statistics and probability*, Ed. Puri, M.L., VSP International Science Publishers, pp. 301-315.

He, G., Müller, H.G. & Wang, J.L. (2003). Functional canonical analysis for square integrable stochastic processes. *J. Multiv. Anal.* **85**, 54-77.

He, G., Müller, H.G., Wang, J.L. (2004). Methods of canonical analysis for functional data. *J. Statist. Plann. and Inf.* **122**, 141-159.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688-698.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statist. Science*, **15**, 1-26.

Heckman, N. & Zamar, R. (2000). Comparing the shapes of regression functions. *Biometrika* **87**, 135-144.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321-377.

James, G. (2002). Generalized linear models with functional predictors. *J. Royal Statist. Soc. B* **64**, 411-432.

James, G., Hastie, T. G. & Sugar, C. A. (2001). Principal component models for sparse functional data. *Biometrika* **87**, 587-602.

James, G. & Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98**, 397-408.

Jones, M.C. & Foster, P.J. (1996). A simple nonnegative boundary correction for kernel density estimation. *Statistica Sinica* **6**, 1005-1013.

Jones, M.C. and Rice, J. (1992). Displaying the Important Features of Large Collections of Similar Curves. *American Statistician* **46**, 140-145.

Ke, C. and Wang, Y. (2001). Semiparametric Nonlinear Mixed-Effects models and Their applications. *Journal of American Statistist Association* **96**, 1272-1281.

Kirkpatrick, M. & Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms and other infinite-dimensional characters. *J. Math. Biol.*, 27, 429-450.

Kneip, A. & Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.*, **16**, 82-112.

Leurgans, S.E., Moyeed, R.A. & Silverman, B.W. (1993). Canonical correlation analysis when the data are curves. *J. Royal Statist. Soc. Series B* **55**, 725-740.

Leng, X. & Müller, H.G. (2006). Time ordering of gene co-expression. *Biostatistics*

Lin, X., Wang, N., Welsh, A. H. & Carroll, R. J. (2004). Equivalent Kernels of Smoothing Splines in Nonparametric Regression for Clustered/Longitudinal Data. *Biometrika* **91**, 177-193.

Lindstrom, M.J. (1995). Self modeling with random shift and scale parameters and a free knot spline shape function. *Statistics in Medicine* **14**, 2009-2021.

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

Liu, X. & Müller, H.G. (2004). Functional convex averaging and synchronization for time-warped random curves. *J. Amer. Statist. Assoc.* **99**, 687-699.

Malfait, N. & Ramsay, J.O. (2003). The historical functional linear model. *Canadian Journal of Statistics* **31**, 115-128.

Marron, J.S., Müller, H.G., Rice, J., Wang, J.L., Wang, N.Y., Wang Y.D., Davidian, M., Diggle, P., Follmann, D., Louis, T.A., Taylor, J., Zeger, S., Goetghebeur, E., Carroll, R.J. (Discussants) (2004). Discussion of nonparametric and semiparametric regression. *Statistica Sinica* **14**, 615-629.

Morris, J.S. & Carroll, R.J. (2006). Wavelet-based functional mixed models. *J. Roy. Statist. Soc. B* **68**, 179199.

Müller, H.G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics **46**, Springer-Verlag, New York.

Müller, H.G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78**, 521-530.

Müller, H.G. (2005). Functional modelling and classification of longitudinal data (with discussion). *Scandinavian J. Statistics* **32**, 223-246.

Müller, H.G., Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics* **33**, 774-805.

Müller, H.G., Zhang, Y. (2005). Time-varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories. *Biometrics* **61**, 1064-1075.

Nadaraya, E.A. (1964). On estimating regression. *Theory Probability Applications* **9**, 141-142.

Pearson, J. D., Morrell, C. H., Brant, L. J., Landis, P. K. & Fleg, J. L. (1997). Gender differences in a longitudinal study of age associated changes in blood pressure. *Journal of Gerontology: Medical Sciences* **52**, 177-183.

Priestley, M. B. & Chao, M. T. (1972). Non-parametric function fitting. *J. Royal Statist. Soc. B* **34**, 385-392.

Ramsay, J. & Dalzell, C.J. (1991). Some tools for functional data analysis. *J. Royal Statist. Soc. Series B* **53**, 539-572.

Ramsay, J. & Li, X. (1998). Curve registration. *J. Royal Statist. Soc. Series B* **60**, 351–363.

Ramsay, J. & Silverman, B. (2002). *Applied functional data analysis*, New York: Springer.

Ramsay, J. & Silverman, B. (2005). *Functional data analysis*, New York: Springer.

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

Rice, J. (2004). Functional and longitudinal data analysis: Perspectives on smoothing. *Statistica Sinica* **14**, 631-647.

Rice, J. & Silverman, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Royal Statist. Soc. Series B*, **53**, 233-243.

Rice, J. & Wu, C. (2000). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253-259.

Rønn, B.B. (2001). Nonparametric maximum likelihood estimation for shifted curves. *J. Royal Statist. Soc. Series B*, **63**, 243-259.

Seifert, B. & Gasser, T. (1996). Finite sample variance of local polynomials: Analysis and solutions. *J. American Statistical Association* **91**, 267-275.

Seifert B. & Gasser, T. (2000). Data adaptive ridging in local polynomial regression. *J. Computational and Graphical Statistics* **9**, 338-360.

Service, S.K., Rice J.A. & Chavez, F.P. (1998). Relationship between physical and biological variables during the upwelling period in Monterey Bay, CA. *Deep-sea research II – topical studies in oceanography* **45**, 1669-1685.

Shi, M., Weiss, R. E. & Taylor, J. M. G. (1996). An analysis of paediatric CD4 counts for Acquired Immune Deficiency Syndrome using flexible random curves. *Applied Statistics*, **45**, 151-163.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Silverman, B.W. (1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society B* **57**, 673–689.

Staniswalis, J.G. & Lee, J.J. (1998). Nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.* **93**, 1403-1418.

Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing.* Chapman & Hall, London.

Wang, N. (2003). Marginal Nonparametric Kernel Regression Accounting for Within-Subject Correlation. *Biometrika* **90**, 43-52.

Wang, Y., Ke, C. & Brown, M.B. (2003). Shape-invariant modeling of circadian rhythms with random effects and smoothing spline ANOVA decompositions. *Biometrics* **59**, 804-812.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā* **A26**, 359-372.

Wu, C. O. & Yu, K. F. (2002). Nonparametric varying coefficient models for the analysis of longitudinal data. *International Statistical Review* **70,** 373-393.

Yao, F., and Lee, T. C. M. (2006). Penalized spline models for functional principal component analysis. *J. Royal Statistical Society B* **68**, 325.

Yao, F., Müller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A. & Vogel, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics* 59, 676-685.

Yao, F., Müller, H.G. & Wang, J.L. (2005a). Functional data analysis for sparse longitudinal data. *J. American Statistical Association* **100**, 577-590.

Yao, F., Müller, H.G. & Wang, J.L. (2005b). Functional linear regression analysis for longitudinal data. *Annals of Statistics* **33**, 2873-2903.

Zeger, S. L., Liang, K. Y. & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049-1060.

Zhao, X., Marron, J.S. & Wells, M.T. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica* **14**, 789-808