CHAPTER 1

Functional modeling of longitudinal data

Hans-Georg Müller

1.1 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, or 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, Chapter 4) or on marginal methods such as Generalized Estimating Equations (GEE, Chapter 3). 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), and the detailed discussion of this topic in Chapter 7). 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 (see 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 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), which will be reviewed in Section 1.5, 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) (see Chapters 10 and 11).

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 and longitudinal data analysis have been revisited more recently, see, for example, 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: First, dependency of the repeated measurements coming from the same subject needs to be taken into account (Lin et al., 2004; Wang, 2003) to improve efficiency of this estimation step. Second, a problem with practical impact for the estimation of a meaningful mean trajectory is individual time variation, which often occurs in addition to amplitude variation, due to differences in time dynamics across subjects. We discuss approaches for handling this issue in Section 1.4.

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 are data-adaptive (Section 1.5). Our focus is less on marginal population-average modeling, although we discuss below the difficulties that arise for marginal modeling in the presence of time variation. 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 Section 1.3.

Once a suitable estimate of the covariance surface is available, one can obtain 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, which 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 1.6). The next two sections (Section 1.2 and Section 1.3 provide background on FDA and smoothing methods. Further discussion can be found in Section 1.7.

### 1.2 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 $Y(t)$, $t \in [0,T]$, for a $T > 0$. 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, while for multivariate data, these notions do not play any role. This is illustrated by the fact that one can reorder 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, and reordering the data will disrupt the analysis.

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 the sample average, $\bar{Y}(t) = \frac{1}{N} \sum_{i=1}^{N} Y_i(t)$, $t \in [0,T]$. 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 variation may be present, in addition to amplitude variation, a challenge 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 solving an “inverse problem” which necessitates some form of regularization, a feature typical for many functional techniques. Two main types of regularization have been used: Regularization by an additive penalty term, usually penalizing against non-smooth curve estimates, and often 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 referred to 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 threshold). Due to the difficulties in calibrating the regularization for functional canonical correlation, alternative notions of functional correlation measures (Service, Rice and Chavez, 1998; Heckman and Zamar, 2000; Dubin and Müller, 2005) have been proposed.

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. Another complex case involves the simultaneous presence of functional predictors and functional responses. These models will be discussed in Section 1.6. 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.

The term “functional data” is used here to denote a sample of curves, rather than a single curve, as one may encounter in dose-response analysis or in nonparametric regression. However, in general, the use of the term “functional data” is 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 (producing a so-called “spaghetti plot”). In other situations, when data are irregularly sampled or a derivative is required, as in the modeling of growth curves, a preliminary smoothing or differentiation step may be helpful (examples are provided in Sections 1.5 and 1.6 and in the left panel of Figure 1.1). From such plots, one may discern a general trend in the data, changes in sampling frequencies (for example caused by dropouts) and the shapes of individual trajectories and how much these shapes vary across
subjects. Last, but not least, one may identify subjects with outlying trajectories; these are candidates for further study or removal before proceeding with the analysis.

1.3 Nonparametric regression

1.3.1 Kernel smoothing

Smoothing methods for nonparametric regression are an important ingredient of many functional methods. These 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 functional 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 (Eubank, 1999).

The goal of smoothing in the nonparametric regression setting is to estimate a smooth regression function or surface \( g(x) = E(Y|X = x) \), usually assumed to be twice continuously differentiable. For the random design case, this regression function is characterized by the joint distribution of vectors \((X,Y)\), while for fixed designs the predictor levels \(X_j\), at which responses \(Y_j\) are recorded, are assumed to be non-random (and usually assumed to be generated by a “design density”). The response \(Y\) is univariate, while predictors \(X\) can be univariate or multivariate. Of interest for FDA applications are the cases \(X \in \mathbb{R}\), the case of a one-dimensional regression function, and \(X \in \mathbb{R}^2\), the case of a regression surface. We note that the notations \(X\) and \(Y\) will be used in subsequent sections to denote functional objects rather than vectors or scalars.

To define a kernel smoother for a one-dimensional predictor, given \(n\) data points \(\{(X_j,Y_j)\}_{1 \leq j \leq n}\), we need a bandwidth or window width \(h\) and a one-dimensional kernel function \(K_1\). The bandwidth serves as smoothing parameter and determines the trade-off between variance and bias of the resulting nonparametric regression estimates. The kernel \(K_1\) 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 with domain \([-1,1]\) are rectangular (box) kernels \(K_1(x) = 1/2\) or quadratic (Epanechnikov) kernels \(K_1(x) = \frac{3}{4}(1-x^2)\), which enjoy some optimality properties. Popular non-negative kernels with unbounded domain are Gaussian kernels, which correspond to the standard normal density.

A classic kernel smoothing method primarily aimed at regular designs \(X_j\) is the class of convolution kernel smoothers (Priestley and Chao, 1972; Gasser and Müller, 1984). The smoothing window for estimating at predictor level \(x\) is \([x-h, x+h]\) if a kernel function \(K_1\) with domain \([-1,1]\) is used. Let \(S(j) = (X(j) + X(j-1))/2\), where \(X(j)\) is the \(j\)-th order statistic of the \(X_j\), and let \(Y[j]\) denote the concomitant of \(X(j)\). Convolution type kernel estimators are defined as

\[
\hat{g}_C(x) = \sum_{j=1}^{n} Y[j] \int_{S(j)}^{S(j+1)} \frac{1}{h} K_1(\frac{x-s}{h}) \, ds. 
\]  

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).
1.3.2 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 and regular 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 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, $\nu$ 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 Figure 1.1.

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 highly 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 were 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 leave-one-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 have been around for a long time and are based on the very simple idea of localizing a linear regression fit from the entire data domain to local windows. 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 the case of 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

$$\sum_{j=1}^{n} K_1 \left( \frac{x - X_j}{b} \right) \left( Y_j - a_0 - a_1(x - X_j) \right)^2 .$$

Once $\hat{a}_0 = \hat{a}_0(x)$ has been obtained, we define the local linear kernel estimator $\hat{g}_L$ as $\hat{g}_L(x) = \hat{a}_0(x)$. The older kernel methods of Nadaraya (1964) and Watson (1964), correspond to the less flexible special case of fitting constants to the data locally by weighted least squares. Fitting local constants however leads to somewhat awkward bias behavior if the designs are non-equidistant. All kernel type estimators as well as smoothing splines exhibit...
very similar behavior in the interior of the domain (away from the boundaries) and when the design is regular. Differences emerge for random designs and when estimating near the boundaries.

For two-dimensional smoothing we aim at the regression function (regression surface) $g(x_1, x_2) = E(Y|X_1 = x_1, X_2 = x_2)$. Locally weighted least squares then provide a criterion for fitting local planes to the data \{$(X_{j1}, X_{j2}, Y_j)_{j=1,...,n}$\}, leading to the surface estimate $\hat{g}(x_1, x_2) = \hat{a}_0$, where $(\hat{a}_0, \hat{a}_1, \hat{a}_2)$ are the minimizers of the locally weighted sum of squares

$$\sum_{j=1}^{M} K_2 \left(\frac{x_1 - X_{j1}}{h_1}, \frac{x_2 - X_{j2}}{h_2}\right) \left[Y_j - \{a_0 + a_1(X_{j1} - x_1) + a_2(X_{j2} - x_2)\}\right]^2.$$  \hspace{1cm} (1.3)

Here, $K_2$ is a two-dimensional kernel function, usually chosen as a two-dimensional density function. A common choice for $K_2$ is the product of two one-dimensional densities, $K_2(u_1, u_2) = K_1(u_1)K_1(u_2)$, where $K_1$ is any of the one-dimensional kernels discussed above. Two bandwidths $h_1, h_2$ are needed; for simplicity they are often chosen to coincide, $h_1 = h_2$, so that only one smoothing parameter needs to be selected. We note that often useful explicit formulas for the smoothing weights employed for both one- and two-dimensional smoothers can be easily obtained (see, e.g., formulas (2.5) in Hall, Müller and Wang, 2006).

1.4 Time warping and curve synchronization

1.4.1 Overview

A main motivation for considering time warping in biomedical applications is the empirical observation that individuals may progress through time at their own individual pace, referred to as biological time or “eigenzeit” (Capra and Müller, 1997). In contrast to clock time, this time is often defined by physiological processes of development, and significant events are defined by reaching certain stages of maturity. These stages are attained earlier for some individuals, and later for others. The idea is then to compare individuals at corresponding stages of their biological development, and not based on chronological age.

A typical example is human growth where various growth spurts have been well identified and provide natural “landmarks” of development. These include the well-known pubertal growth spurt and also the so-called mid-growth spurt that has been rediscovered using non-parametric smoothing methods – see Gasser et al., 1984). As these spurts occur at a different chronological age for each individual, adequate models for corresponding longitudinal data need to reflect the presence of time variation (the variation in the timing of the spurts) in addition to amplitude variation (the variation in the size of the spurts). In the growth curve example, the presence of time variation implies that a simple cross-sectional average growth curve will often not be very meaningful. The reason is that it will not resemble any individual trajectory closely as data obtained for non-corresponding times are averaged, and therefore will lead to wrong impressions about the dynamics of growth. While the phenomenon of time variation is more obvious for some longitudinal data than for others, its presence is always a possibility when dealing with longitudinal data.

Time warping has been studied primarily for densely and regularly sampled data, such as data from longitudinal growth studies, but is of potential relevance for many longitudinal studies. Addressing the warping issue 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 warping models, reflecting the individually determined flow of time, one assumes that the time axis is individually distorted, for example by a random time transformation function that is monotone increasing while beginning and end point of the time domain remain unchanged. We do not attempt to give comprehensive descriptions of the various curve registration methods. The reader is referred
In the presence of warping, the simultaneous random variation in amplitude and time can lead to identifiability issues, and therefore common warping methods often contain implicit assumptions about the nature of the time variation. When each subject follows its own time scale, time synchronization as a pre-processing step often improves subsequent analysis of functional data, and often is a necessity for a meaningful analysis when the time dynamics of longitudinal data is of interest. Analyzing the nature of the time variation often is of interest in itself. For example, in gene time course expression analysis, gene classification can be based on a simple approach of time-shift warping (Silverman, 1995; Leng and Müller, 2006).

A specific example for growth data is shown in Figure 1.1. The left panel displays a sample of growth velocities (obtained by using local polynomial fitting to estimate derivatives) for 54 girls from the Berkeley Growth Study, while the right panel features a comparison of the cross-sectional mean growth curve, displaying various mean functions that were obtained by applying the traditional cross-sectional average as well as various warping (registration) methods. Among these, the landmark method, pioneered in Kneip and Gasser (1992), is known to work very well for these data and serves as benchmark.

In landmark warping one first identifies landmark locations, often defined as peaks or troughs in first or second derivatives of the individual trajectories. These locations are
times of events such as peak growth velocity during a growth spurt that have a meaning-
ful biological interpretation and betray an individual’s time line, for example accelerated
or delayed development. The landmark times are considered to correspond to each other
across individuals, and in the landmark approach the average curve is required to include
both average location and average curve value for each of the landmark events. These char-
acteristic points are supplemented by an interpolation step to define an average smooth
curve which connects the averaged landmarks. In the growth curve example, the landmark
average curve includes the point defined by (i) the time which corresponds to the average
timing of the pubertal growth spurts, defined as the location of maximal growth velocity,
and by (ii) the average value of all maximum growth velocities. It also goes through all
points similarly defined by other landmarks.

The alternative Procrustes method (Ramsay and Li, 1998) is an iterative procedure,
warting curves at each step to match the current cross-sectional mean as closely as possible.
The current cross-sectional mean is then updated and serves again as template for the next
warting step, and this is repeated until convergence. A third method is area under the
curve registration (Liu and Müller, 2004) which synchronizes time points associated with
the same relative area under the curve between the left endpoint of the domain and the
respective time point. When comparing the resulting average growth curves in Figure 1.1
(right panel), the cross-sectional mean is found to underestimate the size of the pubertal
growth spurt and it also produces a biased location. Similar distortions are found for the
midgrowth spur, the smaller growth spur that occurs around 5 years. While the Procrustes
method is improving upon the cross-sectional mean, area under the curve warping is found
to mimic landmark warping the closest in this example. We note that this method is very
simple to implement.

We conclude that even a simple notion such as a mean curve needs to be carefully re-
considered in the presence of time variation. A simulated example further demonstrating
the distorting effects of warping in FDA is shown in Figure 1.2. Here all individual trajec-
tories have been generated as bimodal curves but the cross-sectional mean does not reflect
this shape at all and is clearly inadequate. Modifications aiming at time-synchronization are
needed to arrive at a representative mean when warping is present; the area under the curve
warting method is seen to provide these for this example and nicely recovers the original
shape of individual trajectories.

Landmark identification and alignment (Gasser and Kneip, 1995) has become a gold
standard for warping, for those situations where landmarks are easy to identify (see Figure
1.1). 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. However, landmark
methods 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 shape requirements and are therefore more universally applicable. Alternative
robust warping methods have been developed lately (Gervini and Gasser, 2005). Much work
remains to be done in this area.

1.4.2 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 \(Y_i(t) = Y_0(t - \tau_i)\), \(\tau_i\) denoting a (random) time shift for the \(i\)-th subject, and \(Y_0\) a synchronized tra-
cjectory. 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,
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” \( S \) (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

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

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

The identifiability problem corresponds to the fact that this mapping does not have a unique inverse. The way this ambiguity is resolved differentiates between the various warping methods such as Procrustes method or landmark warping, each providing a concrete way to define an inverse mapping and thus defining a synchronization algorithm.

A simple and often effective warping method that is featured in Figures 1.1 and 1.2 is area under the curve warping. This method is designed for samples of non-negative random trajectories. 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 therefore are mapped to the same synchronized time. Formally, to obtain the inverse warping process $\tilde{X}^{-1}(x)$, 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 as 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 $\psi^{-1}$ 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 $J$ functions, $J > 2$,

$$\bigoplus_{j=1}^J \pi_j Y_j = \psi\left( \sum_{j=1}^J \pi_j X_j + \sum_{j=1}^J \pi_j Y_j \right),$$

for any $\pi_j$ such that $\sum_{j=1}^J \pi_j = 1, 0 \leq \pi_j \leq 1, j = 1, \ldots, J$. This then leads to the warped average function (functional convex average, shown in the right panels of Figures 1.1 and 1.2),

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

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

1.5 Functional principal component analysis

1.5.1 Square integrable stochastic processes

Functional Principal Component Analysis (FPCA) has emerged as a major tool for dimension reduction within FDA. One goal 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

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

$$G(s,t) = \text{cov}\{Y(s),Y(t)\}, \quad s, t \in [0,T]$$

and the auto-covariance operator

$$(Af)(t) = \int_0^T f(s)G(s,t) \, ds.$$
FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

This is a linear Hilbert-Schmidt operator in the function space of square integrable functions $L^2([0, T])$ with Hilbert-Schmidt kernel $G$ (Conway, 1985). Under minimal assumptions this operator has orthonormal eigenfunctions $\psi_k$, $k = 1, 2, \ldots$ with associated ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$, i.e., satisfying

$$A \psi_k = \lambda_k \psi_k, \quad k = 1, 2, \ldots.$$  

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 $Y$,

$$Y(t) = \mu(t) + \sum_{k=1}^{\infty} A_k \psi_k(t), \quad (1.6)$$

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

$$A_k = \int_0^T \{Y(t) - \mu(t)\} \psi_k(t) \, dt. \quad (1.7)$$

The situation is analogous to the representation of random vectors in multivariate analysis by principal components, replacing the inner product in the vector space $\mathbb{R}^d$, 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. Then a random vector can be represented in the basis defined by the eigenvectors of its covariance matrix, which is the finite-dimensional equivalent of the Karhunen-Loève expansion.

1.5.2 From Karhunen-Loève representation to functional principal components

One of the main attractions of FPCA is that knowledge of the distribution of the (uncorrelated) set of FPC scores $\{A_1, A_2, \ldots\}$ is equivalent to knowledge of the distribution of $Y - \mu$, which is a consequence of the Karhunen-Loève expansion. While this equivalence is of theoretical interest, in practice one needs to truncate the sequence of FPC scores at a suitable index (chosen data-adaptively whenever possible). This truncation then corresponds to 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 non-smooth trajectories such as functions containing small jumps and sharp edges. They are less well suited to reproduce smooth trajectories. The disadvantage of fixed basis functions is that they may not be very parsimonious and a larger number of basis functions may be 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) in the context of growth curves. Other key references are Castro, Lawton and Sylvester (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 \([0, T]\) (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. The next step is then 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 focuses on asymptotics as \( N \to \infty \). However, if data are either irregularly or not densely 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 noise-free trajectories is mainly of theoretical interest.

1.5.3 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 in the case of longitudinal data, which from now on we assume to consist of sparse, irregular and noisy measurements of the longitudinal trajectories. According to (1.7), the FPC scores, which are the random effects in the representation, would normally 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 for the case of dense measurements. 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

\[
Y_{ij} = Y_i(t_{ij}) + e_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} A_{ik} \psi_k(t_{ij}) + e_{ij}.
\]

Here the \( e_{ij} \) are i.i.d. measurement errors with moments \( E(e_{ij}) = 0 \), \( E(e_{ij}^2) = \sigma^2 \), and the \( e_{ij} \) 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 (Figure 1.3, left panel).

1.5.4 Principal analysis by conditional expectation

Here we describe the Principal Analysis via Conditional Expectation (PACE) 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 Principal Analysis of Random Trajectories (PART) algorithm for obtaining the empirical Karhunen-Loève representation.
of smooth functional data, where measurements are contaminated with additional measurement error. This algorithm works irrespective of whether the measurements have been sampled on 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).

The PART algorithm consists of the following steps: In a first step, one pools all available measurements \((t_{ij}, Y_{ij})\), \(i = 1, \ldots, N, j = 1, \ldots, n_i\), 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 \((t_{ij}, t_{il})\). These data are then entered into a 2-dimensional scatterplot smoother; for example, the local linear weighted least squares approach (1.3). The output is the estimated covariance surface. The diagonal elements (for which \(j = l\)) are omitted from the input for the 2-dimensional smoothing step, 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 may obtain the variance function of the errors by smoothing along the direction of the diagonal.

While in our implementation we use the local linear smoothers described in Section 1.3,
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, simply by 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. If the observations are noisy or sparse, the Riemann sums will not provide reasonable approximations to the integrals (1.7), and this is where the PACE approach to predict individual FPC scores comes in. For this approach, we make Gaussian assumptions, i.e., $A_{ik}, e_{ij}$ are assumed to be jointly normal. Define the vectors

$$Y_i = (Y_{i1}, \ldots, Y_{in})', \mu_i = \{\mu(t_{i1}), \ldots, \mu(t_{in})\}', \psi_{ik} = \{\psi_k(t_{i1}), \ldots, \psi_k(t_{in})\}'. $$

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

$$E(A_{ik}|Y_i) = E(A_{ik}) + \text{cov}(A_{ik}, Y_i)\text{cov}(Y_i, Y_i)^{-1}(Y_i - \mu_i),$$

where

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

Substituting the estimates discussed above then leads to estimated predicted FPC scores

$$\hat{E}(A_{ik}|Y_i) = \hat{\lambda}_k \hat{\psi}_{ik}' \hat{\Sigma}_{Y_i}^{-1}(Y_i - \hat{\mu}_i).$$

(1.8)

1.5.5 Predicting individual trajectories

Once the number of included random coefficients $K$ has been determined, we can use the predicted FPC scores (1.8) to obtain predicted individual trajectories

$$\hat{Y}(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{E}(A_k|\hat{Y}_i) \hat{\psi}_k(t).$$

(1.9)

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, several options are available. A simple and fast method is the scree plot. Here one plots the fraction of variance unexplained by the first $K$ components against the number of included components $K$,

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

where $K\infty$ is a large number of components, which is clearly larger than the number to be selected. One looks for a “elbow” in this graph, i.e., a value of $K$ at which the rate of decline slows substantially, as $K$ increases further, and the number $K$ where this elbow occurs is the selected number.

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

\[
\hat{L}_1 = \sum_{i=1}^{N} \{- \frac{n_i}{2} \log (2\pi) - \frac{1}{2} \log (\det \hat{\Sigma}_{Y_i}) - \frac{1}{2} (\hat{Y}_i - \hat{\mu}_i)' \hat{\Sigma}_{Y_i}^{-1} (\hat{Y}_i - \hat{\mu}_i)\},
\]

(1.10)

while a conditional version of the likelihood, conditioning on predicted FPC scores, is given by

\[
\hat{L} = \sum_{i=1}^{N} \{- \frac{n_i}{2} \log (2\pi) - \frac{n_i}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} (Y_i - \hat{\mu}_i - \sum_{k=1}^{K} \hat{A}_{ik} \hat{\psi}_{ik})'(\hat{Y}_i - \hat{\mu}_i - \sum_{k=1}^{K} \hat{A}_{ik} \hat{\psi}_{ik})\}.
\]

(1.11)

In either version, the pseudo-AIC value is then \( \text{AIC} = -2\hat{L} + 2K \), and the minimizing value for \( K \) is selected.

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 smoothing it individually, without regard for the other curves in the sample. We note that this traditional approach has proven useful for regular designs as encountered in growth studies, and remains recommended as an exploratory tool. However, this approach ignores the information available across the sample and is thus less efficient than the functional approaches. It is also infeasible for the commonly encountered longitudinal data where the number of observations per curve is small and the locations of the measurements are irregular. In theoretical analysis this situation 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 individual FPC scores have been obtained, 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, and is a very useful exploratory tool, for example to identify clusters in the data. Pairs or vectors of FPC scores have been successfully employed for more formal 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 (an early paper is Dauxois, Pousse and Romain, 1982). Assuming more than one but at most finitely many observations are available per trajectory, and without making Gaussian assumptions, it was shown in Hall et al. (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 and 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 as obtained from the PACE method converge to the actual FPC scores as the designs get asymptotically denser (more and more measurements per subject, see Müller, 2005). Under additional Gaussian assumptions, 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).
1.5.6 Application to longitudinal CD4 data

As an illustration, the PACE method was applied to longitudinal 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. These data fit the description of sparse and irregular data and are shown in the left panel of Figure 1.3, 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 Figure 1.3, where the diagonal has been omitted as described above. The overall mean function $\hat{\mu}(t)$ is depicted in the lower panels of Figure 1.4. The conditional pseudo-likelihood (1.11) 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 setting 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 Figure 1.4 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 Figure 1.4; 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. Nevertheless, as this example shows, an added benefit of eigenfunctions is their interpretability as (orthogonal) modes of variation that in their entirety explain the observed variation in the functional data. The shapes of individual eigenfunctions often associate a concrete description with these modes of variation.

Finally, we are interested in modeling individual trajectories, which are obtained via the estimated FPC scores, see equation (1.9). The predicted trajectories for four subjects, including confidence bands, are shown in Figure 1.5, 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.

1.6 Functional regression models

1.6.1 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 may include a random trajectory in either predictors, responses, or both. We assume here that 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 $\mu_X, \mu_Y$. In the functional case we denote eigenfunctions by $\phi_k$ for $X$ and $\psi_k$ for $Y$. 
The linear model for a scalar response and a functional predictor is

\[ E(Y|X) = \mu_Y + \int_0^S \{X(s) - \mu_X(s)\} \beta(s) \, ds, \]

where \( \beta \) is the regression parameter function and \([0, S]\) is the domain of \( X \). An extension to functional responses is the model

\[ E(Y(t)|X) = \mu_Y(t) + \int_0^S \{X(s) - \mu_X(s)\} \beta(s, t) \, ds, \]

(1.12)

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
Figure 1.5 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, obtained under Gaussian assumptions. 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 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 \( X \in \mathbb{R}^p, \ Y \in \mathbb{R}^q \), the normal equation is \( \text{cov}(X,Y) = \text{cov}(X)B \), where \( \text{cov}(X,Y) \) is the \( p \times q \) matrix with elements \( a_{jk} = \text{cov}(X_j,Y_k) \). This equation can be solved for \( B \) if the \( p \times p \) covariance matrix \( \text{cov}(X) \) 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
\[ r_{XX}(s, t) = \text{cov}\{X(s), X(t)\}, \quad r_{XY}(s, t) = \text{cov}\{X(s), Y(t)\}. \]

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 et al., 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
\[
E(Y|X) = g\left\{ \mu + \int_0^S X(s)\beta(s)\,ds \right\},
\]
(1.13)
coupled with a variance function \( \text{var}(Y|X) = V\{E(Y|X)\} \). This model is an extension of the common Generalized Linear Model (GLM) to the case of functional predictors. 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,
\[
E\{Y(t)|Z = z\} = \mu(t) + \sum_{k=1}^K \alpha_k(\gamma_k z)\psi_k(t).
\]
Here the \( \gamma_k \) are single indices (i.e., vectors which project the covariates \( Z \) into one dimension), and the \( \alpha_k \) are link functions to the random effects. Sometimes simpler structured models such as a “multiplicative effects model” are useful,
\[
\mu(t, z) = \mu_0(t)\theta(z), \quad E\{Y(t)\} = \mu_0(t), \quad E\{\theta(Z)\} = 1,
\]
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 (2003) and Müller and Zhang (2005). In the latter paper, the response is remaining lifetime, to be predicted from a longitudinal covariate which is observed up 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
\[
E\{Y(t)|X\} = \mu_Y(t) + \zeta(t)X(t)
\]
(under the assumption of one predictor process) emerges as a special case; here \( \zeta(\cdot) \) 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 demonstrate this feature for the functional regression model (1.12).

### 1.6.2 Functional regression for longitudinal data

Extending the functional linear regression model (1.12) introduced above 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,
\[ U_{it} = X_i(s_{it}) + e_{it} = \mu_X(s_{it}) + \sum_{m=1}^{\infty} A_{im} \phi_m(s_{it}) + e_{it}, \quad s_{it} \in [0, S], \]
\[ V_{ij} = Y_i(t_{ij}) + e_{ij} = \mu_Y(t_{ij}) + \sum_{k=1}^{\infty} B_{ik} \psi_k(t_{ij}) + e_{ij}, \quad t_{ij} \in [0, T]. \]
Here the times \( s_{ij} \), resp. \( t_{ij} \), where measurements are recorded for predictor processes \( X \), resp. 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. Measurements for both processes are not only sparse and irregular, but also contaminated by measurement errors \( e_{it}, e_{ij} \) which are assumed to satisfy the properties listed in Subsection 1.5.3.

Applying FPCA, by using the orthonormality properties of the eigenfunctions, one finds that the regression parameter function \( \beta \) in (1.12) can be represented by
\[ \beta(s, t) = \sum_{k,m=1}^{\infty} \frac{E(A_{m} B_{k})}{E(A_{m}^2)} \phi_m(s) \psi_k(t). \quad (1.14) \]
This reduces the problem to estimate \( \beta \) to the problem to obtain an estimate of \( E(A_{m} B_{k}) \), for which we consider
\[ \widehat{E}(A_{m} B_{k}) = \int_{0}^{T} \int_{0}^{S} \hat{\phi}_m(s) \hat{\Gamma}_{XY}(s,t) \hat{\psi}_k(t) \, ds \, dt, \quad (1.15) \]
where \( \hat{\Gamma}_{XY}(s,t) \) is a local linear smoother for the cross-covariance function \( \Gamma_{XY}(s,t) = \text{cov}(X(s), Y(t)) \) (Yao et al., 2005b), and the integral is evaluated by numerical integration.

Once the regression parameter surface \( \beta \) 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 \( U^* = (U^*_1, \ldots, U^*_L)' \) available for \( X^*(\cdot) \), where \( U^* \) denotes data available on the predictor process for one individual. Under Gaussian assumptions, the best predictor is given by
\[ E\{Y^*(t)|X^*(\cdot)\} = \mu_Y(t) + \int_{0}^{S} \beta(s,t)\{X^*(s) - \mu_X(s)\} \, ds \\
= \mu_Y(t) + \sum_{k,m=1}^{\infty} \frac{E(A_{m} B_{k})}{E(A_{m}^2)} A_{m} \psi_k(t). \]
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 processes \( X \) resp. \( Y \), we arrive at
\[ \hat{Y}_{KM}^*(t) = \hat{\mu}_Y(t) + \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{\widehat{E}(A_{m} B_{k})}{\overline{E}(A_{m}^2)} \widehat{E}(A_{m}^* U^*) \hat{\psi}_k(t), \quad (1.16) \]
where \( \overline{E}(A_{m}^* U^*) \) is estimated by the PACE method, as described in Subsection 1.5.4, given observations \( U^* = (U^*_1, \ldots, U^*_L)' \) 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 \( R^2 = \text{var}\{E(Y|X)\}/\text{var}(Y) \), which is used to measure the strength of a regression relationship, to the functional case. Applying orthonormality properties of the eigenfunctions, one such
possible extension can be represented as

\[
R^2 = \frac{\int_0^T \text{var}[E(Y(t)|X)]dt}{\int_0^T \text{var}(Y(t))dt} = \sum_{k,m=1}^{\infty} \frac{E(A_m^2 B_k^2)/E(A_m^2)}{\sum_{k=1}^{\infty} E(B_k^2)}.
\] (1.17)

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 (1.15). Substituting the corresponding estimates then leads to the estimated functional coefficient of determination \(R^2\).

1.6.3 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 being 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 Figure 1.6. 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.

![Figure 1.6](image-url)

Running the functional regression machinery, we obtain the estimate of the regression surface function \(\hat{\beta}(\cdot, \cdot)\) for these data, as depicted in Figure 1.7. 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 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 later 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.

![Figure 1.7 Estimated regression parameter surface $\beta (1.14)$ 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.](image)

Finally, predicted trajectories of systolic blood pressure for four randomly selected participants are displayed in Figure 1.8. 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, however the uncorrelatedness feature of the predictors will be lost in this case.

### 1.7 Concluding remarks and outlook

Functional data analysis (FDA) 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 with a focus on modeling and data exploration under very flexible model assumptions with no or few parametric components. In this chapter, we reviewed some of the tools of functional data analysis which include smoothing, functional principal components, functional linear models and time-warping. Warping or curve registration is a FDA-specific methodology that aims at adjusting for random time distortions.
While in the usual functional data analysis paradigm the sample functions are considered as continuously observed, in longitudinal data analysis one mostly deals with sparsely and irregularly observed data that also are corrupted with noise. We described some of the adjustments of the FDA techniques which are needed to take full advantage of the FDA approach when analyzing longitudinal data. The extension of FDA towards the analysis of longitudinal data is a fairly recent undertaking that presents a promising avenue for future research.

In addition to the FDA methodology described in this chapter, several other FDA approaches have applications for longitudinal data. These include functional ANOVA decompositions using smoothing spline models, proposed in Brumback and Rice (1998) and applications of P-splines (Bugli and Lambert, 2006) (see Chapters 10 and 11). Another class of non- or semiparametric models that is of interest for longitudinal studies are varying coefficient models where \( Y(t) \) is related to a series of predictors \( X_1(t), \ldots, X_p(t) \). In a typical two-step procedure (Fan and Zhang, 2000), one conducts a linear regression at each time point that is an element of a grid of time points and then one applies smoothing to the resulting regression coefficients to obtain smooth varying-coefficient functions. In addition, shape-invariant modeling is a promising functional method, as it combines the analysis of time variation (warping) with that of amplitude variation.

For functional inference, bootstrap methods based on resampling subjects, keeping data together that belong to the same subject, have been used, and are very promising, for both construction of confidence regions and for significance tests. Asymptotic inference is avail-
able under Gaussian assumptions, but is not yet available on a wider scale; compare Fan and Lin (1998). This chapter demonstrates that functional approaches provide a flexible alternative to common parametric models for analyzing longitudinal data. Software is currently available in the form of specialized R and Matlab programs offered by various researchers and is under development. No definitive packages have emerged as of yet.

At this time, a number of key techniques are in place, notably smoothing and differentiation of noisy data, warping and curve registration, functional principal component analysis, functional regression and penalized regularization techniques. The unique combination of methods from functional analysis, stochastic processes, multivariate analysis, smoothing and the many open questions make the interface of FDA and longitudinal data analysis a rewarding research area. Practitioners can benefit from the highly flexible FDA toolbox which is useful for both exploratory analysis and inference and facilitates new insights into the dynamics of longitudinal data.

Acknowledgments

I am indebted to two editorial reviewers for very careful reading and numerous suggestions which led to improvements. This work was supported in part by National Science Foundation grants DMS03-54448 and DMS05-05537.
References


REFERENCES


