Functional Modeling and Classification of Longitudinal Data

HANS-GEORG MÜLLER

University of California, Davis

Running head: Functional Modeling

ABSTRACT. We review and extend some statistical tools that have proved useful for analyzing functional data. Functional data analysis primarily is designed for the analysis of random trajectories and infinite-dimensional data, and there exists a need for the development of adequate statistical estimation and inference techniques. While this field is in flux, some methods have proven useful. These include warping methods, functional principal component analysis, and conditioning under Gaussian assumptions for the case of sparse data. The latter is a recent development that may provide a bridge between functional and more classical longitudinal data analysis. Besides presenting a brief review of functional principal components and functional regression, we develop some concepts for estimating functional principal component scores in the sparse situation. An extension of the so-called generalized functional linear model to the case of sparse longitudinal predictors is proposed. This extension includes functional binary regression models for longitudinal data and is illustrated with data on primary biliary cirrhosis.

Key words: Binary regression, discriminant analysis, functional data analysis, generalized functional linear model, logistic model, longitudinal data, principal components, sparseness, stochastic process.

Correspondence: Hans-Georg Müller, *Email*: mueller@wald.ucdavis.edu, *Address*: Department of Statistics, One Shields Ave., University of California, Davis, CA 95616, USA.

1. Introduction

While Multivariate Data Analysis (MDA) is concerned with data in the form of random vectors, Functional Data Analysis (FDA) goes one big step further, focusing on data that are infinite-dimensional, such as curves, shapes and images. Data of this type are increasingly collected, due to technological advances in measurement devices and computing. Another driving force is that increasingly, complex scientific questions are being pursued, where time dynamics or spatial dynamics are a major component. Extending one-dimensional to multivariate objects is the focus of MDA, but this step cannot be simply extrapolated to extend the vector case to the functional, infinite-dimensional case. In many ways the step from finite to infinite dimension is a bigger leap than the step from one to finite dimensions, at least from a conceptual and theoretical point of view. On the other hand, proven methods of multivariate analysis form the backbone for some of the prominent techniques of FDA. An overarching theme in FDA is the necessity to achieve some form of dimension reduction of the infinite-dimensional data to finite and tractable dimensions. For theoretical analysis, dimensions are usually assumed to increase with increasing sample size, necessitating the development of increasing-dimension asymptotics, which poses interesting theoretical challenges. Areas that FDA draws upon besides multivariate analysis include smoothing, especially nonparametric regression, functional analysis (linear operators in Hilbert space), and properties of square integrable stochastic processes.

For an introduction to the field of FDA, the two monographs by Ramsay & Silverman (1997, 2002) provide a rewarding and accessible overview on foundations and applications, as well as a plethora of motivating examples. A special issue of *Statistica Sinica* appeared in July 2004 that includes a focus on recent developments in FDA, with an excellent review article by Rice (2004). Historically, the field was pioneered by Karhunen (1946), who developed foundational theory on stochastic processes in Hilbert space, and by Grenander (1950) with the first systematic application of the Karhunen-Loève expansion to functional data, including the first proposal for functional regression. C.R. Rao (1958) developed preliminary ideas on functional principal components in applications to growth curves. Other notable developments have been a systematic study of the functional analysis that underpins FDA by the Toulouse school of FDA (Dauxois *et al.*, 1982). On the more applied side, the importance of smoothing methods for FDA, including the estimation of derivatives, was demonstrated in Gasser *et al.*(1984), and their importance for functional principal component analysis was pointed out by Rice & Silverman (1991). Functional regression models became popular after their discussion by Ramsay & Dalzell (1992).

A specific aspect of functional data is the possibility of random time transformations. The corresponding warping (curve registration, alignment) procedures have emerged as important tools for some applications, ranging from speech recognition to biological growth and gene expression profiles. When curve data are recorded, different subjects or experimental units may experience events at a

subject-specific pace. A typical example is the human growth curve (Gasser *et al.*, 1984; Kneip & Gasser, 1992), where events such as the pubertal growth spurt occur at different times for different individuals. A large fraction of the variability in a sample of curve data is then best explained as time variation. Landmark method (Gasser & Kneip, 1995), Procrustes method (Ramsay & Li, 1998), time acceleration models (Capra & Müller, 1997), maximum likelihood based alignment (Rønn, 2001) and other forms of curve registration (Wang & Gasser, 1997,1998,1999; Kneip *et al.*, 2000), and recently time-synchronization of random processes (Liu & Müller, 2004) have been proposed, and this subfield is still in rapid development.

This article contains a review of some basic FDA tools such as functional principal component analysis. The topics selected for this review are a subjective selection of a few basic ideas of current developments in FDA and no attempt is made at being comprehensive. In addition to the review, this article contains a proposal to extend generalized functional linear models to the case of sparse and irregular data, with an emphasis on functional binary regression and functional logistic discrimination for the classification of longitudinal data. This proposed methodology is illustrated with longitudinal data on primary biliary cirrhosis patients, with the aim to classify early time courses of bilirubin, observed under sparse measurements, in regard to the long-term survival prospects of the subjects. We consider the case where per subject or experimental unit, one samples one or several functions $X(t)$, $t \in I$ where $I \subset \Re$ is a domain, usually an interval. The observed trajectories typically correspond to a sample of densely sampled longitudinal data. These data are viewed as independent realizations of a stochastic process with smooth trajectories.

Numerous approaches to analyze data in the form of curves have been proposed, including nonlinear parametric regression models (reviewed in Davidian & Giltinan, 1993), state space modeling (reviewed in Jones, 1993), shape-invariant modeling (Stützle *et al.*, 1980) and varying-coefficient models (Fan & Zhang, 1998). The FDA approach is inherently nonparametric, as it lets "the data speak for themselves" by avoiding parametric assumptions, so as not to prejudice the outcome and to enable flexible modeling within a statistical framework that allows to obtain (asymptotic) inference (compare Fan & Lin, 1998). An example that supports this principle is the pre-pubertal growth spurt that had almost vanished from pediatrics textbooks of the 1980's, since the parametric models that had become popular for fitting the human growth curve did not have room for such a second growth spurt – it took a nonparametric analysis to bring this growth spurt (which had been recognized in the pre-modeling era) back on the map (Gasser *et al.*, 1985).

Of current interest in FDA is the extendability of the FDA methodology to the case of longitudinal data. Such data are ubiquitous and there exists a strong need in the life sciences, social sciences and other fields for appropriate flexible methodology. It was highlighted in a recent discussion (Marron

et al., 2004, p. 619) that unifying theory for functional and longitudinal data analysis is one of the pressing open problems in FDA. In many longitudinal studies the data are contaminated by noise and the available repeated measurements for individuals are obtained on sparse and irregular time grids, rather than on dense and regular time grids, as are normally assumed for FDA. Analyzing sparse situations thus is of great practical interest but poses practical and theoretical challenges. Functional methods provide a variety of valuable and potentially powerful tools for longitudinal data analysis if a bridge can be built in which longitudinal data, sampled on sparse designs, can be brought under the umbrella of functional tools. In the framework of functional random effects models, sparse data can be handled by incorporating random effects into models based on B-splines (Shi *et al.*, 1996; Rice & Wu, 2000; James *et al.*, 2001; James & Sugar, 2003). A nonparametric attempt at predicting functional principal component scores from sparsely observed trajectories, including inference through confidence bands, has recently been made in Yao *et al.*(2005a). The connections between FDA and longitudinal data analysis will be explored further below.

If we adopt the nonparametric FDA point of view, basic statistical notions such as mean, variance, correlation and regression need to be developed from scratch again, as the classical notions do not lend themselves to straightforward extensions to the functional case. This is due to basic differences between FDA and MDA: Not only are data in FDA of extremely high dimension, but also affected by time-neighborhood and smoothness relations; time-order is crucial. The analysis changes in a basic way whenever the time order of observations is changed. In contrast, in multivariate statistical analysis, the order of the components of observed random vectors is quite irrelevant, and any change in this order leads to the same results. This fact and the continuous flow of time which serves as argument lead to differences in perspective: Smoothing methods become an indispensable tool in FDA, and the role of the time axis opens the door to warping as an additional dimension of variation.

A basic component of the FDA toolbox is functional principal component analysis (PCA), inherited from multivariate PCA, but with sufficiently different features to merit extra study. The review part of this article is focused on functional PCA (section 2) and on functional regression, which to a large extent uses functional PCA (section 3). No attempt is made at a comprehensive survey of the literature in these areas or of FDA in general. The perspective of extending functional principal components methodology to sparse and irregular longitudinal data will be discussed throughout. In addition to the review, section 3 also contains the proposal to extend the generalized functional linear model to the case of longitudinal, i.e., noisy, sparse and irregular data. The application of this technique and of functional binomial regression to the classification of longitudinal data on primary biliary cirrhosis is reported in section 4. Concluding remarks are in section 5.

2. Functional Principal Components Analysis for functional and longitudinal data

Principal component analysis (PCA) has become a major tool of multivariate analysis since its introduction by Hotelling (1933); compare Joliffe (2002). This method can be extended directly from MDA to FDA. The principle of this extension is to replace vectors by functions, matrices by linear operators, and in particular covariance matrices by auto-covariance operators; scalar products in vector space are replaced by scalar products in function space, which is usually chosen as the space of square integrable functions L^2 on a suitable domain. The principal component functions or eigenfunctions in FDA describe the major "Modes of Variation" of the data (Castro *et al.*, 1986). The importance of smoothing in the estimation of functional principal components has been emphasized in Rice & Silverman (1991), Capra & Müller (1997), Boente & Fraiman (2000) and Cardot (2000). Implementations and illustrations of functional PCA include Kirkpatrick & Heckman (1989), James *et al.*(2001), and Ramsay & Silverman (2002).

Assume the random trajectories X of an underlying stochastic process in $L^2(I)$ have moments as follows: A mean function $\mu(t) = E(X(t))$ and a covariance function $G(s,t) = Cov\{X(s), X(t)\}$, $s,t \in I$. We then define the auto-covariance operator

$$(A_G f)(t) = \int f(s)G(s,t)\,ds$$

which is a linear integral operator with kernel G (compare Courant & Hilbert, 1953). An eigenfunction h of the operator A_G is a solution of the equation $(A_G h)(t) = \lambda h(t)$, with eigenvalue λ. We assume that operators A_G have a sequence of orthonormal eigenfunctions ϕ_k satisfying $\int \phi_j(t)\phi_k(t)\,dt = \delta_{jk}$ (where δ_{jk} is the Kronecker symbol), with ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$, i.e., $(A_G\phi_k)(t) = \lambda_k\,\phi_k(t)$. The kernel G is a Hilbert-Schmidt kernel which can be represented as $G(s,t) = \sum_{k=1}^{\infty} \lambda_k\phi_k(s)\phi_k(t)$, and the underlying process can be represented through the Karhunen-Loève expansion (Karhunen, 1943)

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k\,\phi_k(t), \tag{1}$$

where the sum is defined in the sense of L^2 convergence, with uniform convergence a consequence of Mercer's theorem, and the expansion coefficients ξ_k are uncorrelated random variables (r.v.s), with $E(\xi_k) = 0$ and $\mathrm{var}(\xi_k) = \lambda_k$, such that $\sum_k \lambda_k < \infty$ and

$$\xi_k = \langle X - \mu,\, \phi_k \rangle = \int (X(t) - \mu(t))\phi_k(t)\,dt. \tag{2}$$

The ξ_k are frequently referred to as functional principal component scores and act as random effects in statistical models with functional principal components. Functional PCA, just as regular PCA, can be interpreted as an expansion of the random trajectories X in a functional basis, the eigenfunction

basis of the auto-covariance operator which is an orthonormal basis of the Hilbert space $L^2(I)$. In multivariate analysis the eigenvectors form the corresponding basis of a finite-dimensional vector space. The expansion into eigenfunctions provides a "sparse" representation in the sense that for any finite K, the "fraction of variance explained" by the truncated expansion $X(t) = \mu(t) + \sum_{k=1}^{K} \tilde{\xi}_k \, \phi_k(t)$, which is $F(K) = \{\sum_{k=1}^{K} \lambda_k\}/\{\sum_{k=1}^{\infty} \lambda_k\}$, is maximized among all expansions of X into K basis functions.

The implementation of functional principal components requires some form of regularization (Rice & Silverman, 1991) which can be achieved with smoothing methods; smoothing splines, B-splines, P-splines (compare Ruppert *et al.*, 2003), kernel smoothing and local weighted least squares have been considered for this purpose. Nonparametric implementations with kernel-type smoothing methods were developed in Capra & Müller (1997) and Staniswalis & Lee (1998). Briefly, one estimates the mean function by passing a scatterplot smoother through the aggregated data (t_{ij}, Y_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, where t_{ij}, Y_{ij} denote time, respectively, value of the j-th measurement on the i-th subject. In a model without additional noise, $Y_{ij} = X(t_{ij})$. After obtaining an estimated mean function $\mu(t)$, one may compute raw covariances from all observed pairs of data points for the same subject, $(t_{ij}, Y_{ij}), (t_{il}, Y_{il})$, by $G_{ijl} = (Y_{ij} - \widehat{\mu}(t_{ij}))(Y_{il} - \widehat{\mu}(t_{il}))$. A second surface smoothing step is then applied to the scatterplot $((t_{ij}, t_{il}), G_{ijl})$ in order to obtain the estimated covariance surface.

These smoothing steps can be easily implemented with locally weighted least squares. This smoothing method has several advantages over other regularization methods, as it is straightforward to implement, draws on familiar regression strategies, and the smoothing can be controlled in a straightforward manner. Equally important is that this smoother, along with kernel methods, is based on explicit smoothing weights and is therefore mathematically tractable, and both finite and asymptotic properties have been extensively studied. This is of particular value in FDA, where theory is generally less straightforward to develop and a mathematically tractable smoothing method is a prerequisite for developing asymptotics. We use the local linear scatterplot smoother $S(t)$ at point t with bandwidth h for a scatterplot (V_j, W_j), $j = 1, \ldots, n$, which is the minimizer of $\sum_{j=1}^{n} \kappa(\{V_i - t\}/h)\{W_i - \beta_0 - \beta_1(t - V_i)\}^2$ with respect to β_0, β_1, where a univariate density κ serves as kernel function. Then $S(t) = \hat{\beta}_0(t)$, for which one has an explicit representation that is linear in the W_j (Fan & Gijbels, 1996). Analogously, surface data are smoothed by fitting local planes by weighted least squares.

Applying these or other smoothing procedures, one obtains consistent estimates for mean function μ and covariance function G, and the smoothed covariance function is then discretized on a suitable finite grid, on which it is represented as a covariance matrix. The corresponding spectral decomposition of this matrix yields eigenvectors and eigenvalues. The eigenvectors may be subjected to a final smoothing step to obtain estimated eigenfunctions (Capra & Müller, 1997). More direct procedures without smoothing are possible for the case where one observes entire trajectories on the continuum (Dauxois *et al.*, 1982;

Bosq, 1991; Cardot *et al.*, 2003). Once estimates $\hat{\phi}_k, \hat{\lambda}_k$ of eigenfunctions ϕ_k and eigenvalues λ_k have been obtained, the fitting of individual trajectories requires estimation of functional principal component scores. These can be obtained through the definition $\xi_{ik} = \int (X_i(t) - \mu(t))\phi_k(t)dt$, by plugging estimated mean and eigenfunctions into a Riemann sum approximation of the integral,

$$\hat{\xi}_{ik} = \sum_{j=1}^{N_i} (Y_{ij} - \hat{\mu}(t_{ij}))\hat{\phi}_k(t_{ij})(t_{ij} - t_{i,j-1}), \tag{3}$$

setting $t_{i0} = 0$. This works reasonably well if the support points t_{ij} are densely spaced and the measurements are made without error, so that they reflect the trajectories closely. This approximation works less well if the measurements carry additional measurement errors as is frequently the case (Yao *et al.*, 2003). Additional shrinkage will then improve the estimates.

For the case of noisy, sparse and irregular functional data, as frequently observed in longitudinal studies, the above sum approximation to the integral is clearly doomed. A fully nonparametric approach denoted as PACE (Principal Analysis through Conditional Expectation) to handle such situations was developed in Yao *et al.*(2005a). An alternative practical approach based on B-splines was developed in James *et al.*(2001) and James & Sugar (2003). The sparseness of the data can be modeled by assuming that a random number N_i of measurements T_{ij} is available for the i-th subject such that the T_{ij} are i.i.d. r.v.s. Including additional measurement errors, this leads to a model with measurements U_{ij} made at T_{ij} according to

$$U_{ij} = X_i(T_{ij}) + \epsilon_{ij} = \mu(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(T_{ij}) + \epsilon_{ij}, \tag{4}$$

where the measurement errors ϵ_{ij} are i.i.d. and independent of all other r.v.s, with $E\epsilon_{ij} = 0$, $\mathrm{var}(\epsilon_{ij}) = \sigma^2$.

We then follow through with the same smoothing steps as described above, i.e., estimate the mean function $\mu(t)$ by applying a local linear smoother to the scatterplot $\{(T_{ij}, U_{ij}) : 1 \le i \le n, 1 \le j \le N_i\}$, and the covariance surface $G(s,t)$ by first defining raw estimates $\mathrm{G}_i(T_{ij}, T_{il}) = (U_{ij} - \hat{\mu}(T_{ij}))(U_{il} - \hat{\mu}(T_{il}))$, $j \neq l$ and then applying two-dimensional smoothing to the scatterplot $\{(T_{ij}, T_{il}); G_i(T_{ij}, T_{il})) : 1 \le i \le n, 1 \le j \neq l \le N_i\}$. In this covariance estimation step, the diagonal elements are omitted, as these carry an additional error variance term according to $\mathrm{var}(U_{ij}) = G(T_{ij}, T_{ij})) + \sigma^2$, so that the overall covariance surface has a discontinuity along the diagonal. This phenomenon is illustrated in Fig. 3, and motivates the estimation of σ^2 by decomposing the diagonal of the covariance surface into an estimate as obtained from the smooth surface above, omitting the diagonal elements, and into an estimate exclusively along the diagonal; averaging the difference over a suitable domain then targets σ^2. This idea can be extended in an obvious way to accommodate heteroscedastic errors, characterized by a variance function $\sigma^2(t)$.

The estimation of eigenvalues and eigenfunctions proceeds in the same way as described above for the densely sampled case.

In order to obtain predicted trajectories for sparse, irregular and noisy data, an alternative to the Riemann sum (3) needs to be found. A joint Gaussian distribution assumption for both processes X and errors ϵ_{ij} proves useful. Defining $\widetilde{U}_i = (U_{i1}, \cdots, U_{iN_i})^T$, $\mu_i = (\mu(T_{i1}), \cdots, \mu(T_{iN_i}))^T$ and $\phi_{ik} = (\phi_k(T_{i1}), \cdots, \phi_k(T_{iN_i}))^T$, we obtain the best linear predictors for the k-th functional principal component score for the i-th subject ξ_{ik}, given the sparse noisy observations $\widetilde{U}_i$, as conditional expectation $E[\xi_{ik}|\widetilde{U}_i]$, which under Gaussian assumptions can be explicitly calculated,

$$E[\xi_{ik}|\widetilde{U}_i] = E(\xi_{ik}) + \mathrm{cov}(\xi_{ik}, \widetilde{U}_i)\mathrm{cov}(\widetilde{U}_i, \widetilde{U}_i)^{-1}(\widetilde{U}_i - \mu_i) = \lambda_k \phi_{ik}^T \Sigma_{U_i}^{-1}(\widetilde{U}_i - \mu_i), \tag{5}$$

where

$$(\Sigma_{U_i})_{j,l} = \mathrm{cov}(U_{ij}, U_{il}) = \mathrm{cov}(X_i(T_{ij}), X_i(T_{il})) + \sigma^2 \delta_{jl} = G(T_{ij}, T_{il}) + \sigma^2 \delta_{jl},$$

δ_{jl} as before denoting the Kronecker symbol (see Theorem 3.2.4 in Mardia*et al.*, 1979). Plugging in estimates of $\sigma^2, \lambda_k, \phi_k$ and G then yields the estimated predicted functional principal component score

$$\hat{\xi}_{ik} = \widehat{E}[\xi_{ik}|\widetilde{U}_i] = \hat{\lambda}_k \hat{\phi}_{ik}^T \hat{\Sigma}_{U_i}^{-1}(\widetilde{U}_i - \hat{\mu}_i). \tag{6}$$

This method is easy to implement, effective and has reasonably good asymptotic properties.

One can extend this approach to any linear functional of the random process X. Write such a functional in the form $\langle X - \mu, \zeta \rangle = \int (X(t) - \mu(t))\zeta(t)\, dt$ for a suitable function ζ. Then, with $\zeta_k = \int \zeta(t)\phi_k(t)\, dt$, we have

$$\langle X - \mu, \zeta \rangle = \int (X(t) - \mu(t))\zeta(t)\, dt \; = \sum_{k=1}^{\infty} \xi_k \zeta_k, \tag{7}$$

and therefore, omitting the subject index i, $E[\langle X - \mu, \zeta \rangle | \widetilde{U}] = \sum_k E[\xi_k|\widetilde{U}]\zeta_k$, leading via (6) to estimated best linear predictions

$$\widehat{E}[\int (X(t) - \mu(t))\zeta(t)\, dt | \widetilde{U}] = \sum_k [\hat{\lambda}_k \hat{\phi}_k^T \hat{\Sigma}_U^{-1}(\widetilde{U} - \hat{\mu})]\zeta_k. \tag{8}$$

The following heuristic illustrates the transition from the sparse to the dense case: For simplicity, we consider the underlying (not estimated) quantities as in (5) and the case where the observations are not contaminated by noise. Then, for a fixed t, $\mathrm{cov}(\xi_k, X(t)) = \lambda_k \phi_k(t)$ and

$$E(\xi_k|X(t)) = \frac{\mathrm{cov}(\xi_k, X(t))}{\mathrm{var}(X(t))}[(X(t) - \mu(t)].$$

Extending this to the m-dimensional case, writing for any $t_1, \ldots, t_m$, $E(\xi_k|X(t_1), \ldots, X(t_m)) = \sum_{l=1}^m \beta_{lk}(X(t_l) - \mu(t_l))$, we analyze the behaviour of the β_{lk} for large m. Assume $m \to \infty$ and that

the t_p, $1 \le p \le m$ equidistantly fill out the domain of X. Approximating the covariance matrix of $(X(t_1), \ldots, X(t_m))^T$ by using truncated covariances $\widetilde{\mathrm{cov}}(X(s), X(t)) = \sum_{j=1}^{m} \lambda_j \phi_j(t)\phi_j(s)$ and approximate eigenvectors Δe_j, where $e_j = (\phi_j(t_1), \ldots, \phi_j(t_m))$ and $\Delta = |I|/m$, I denoting the domain of X, we obtain approximate inverse covariance matrices $\widetilde{\mathrm{cov}}^{-1}(X(s), X(t)) = \Delta^2 \sum_{j=1}^{m} \lambda_j^{-1} e_j e_j^T$. For large m, observing $\Delta \sum_{p=1}^{m} \phi_j(t_p)\phi_k(t_p) \to \delta_{jk}$,

$$\begin{aligned} \beta_{lk} &\approx \sum_{p=1}^{m}\sum_{j=1}^{m} \Delta^2 \lambda_j^{-1} \phi_j(t_l)\phi_j(t_p)\lambda_k\phi_k(t_p) \approx \Delta\lambda_k \sum_{j=1}^{m} \lambda_j^{-1}\phi_j(t_l)\Delta\lambda_j^{-1}\phi_j(t_l)\Delta \sum_{p=1}^{m}\phi_j(t_p)\phi_k(t_p) \\ &\approx \Delta\phi_k(t_l). \end{aligned}$$

It follows that for large m,

$$E(\xi_k | X(t_1), \ldots, X(t_m)) = \sum_{l=1}^{m} \beta_{lk}(X(t_l) - \mu(t_l)) \approx \Delta \sum_{l=1}^{m} \phi_k(t_l)(X(t_l) - \mu(t_l)) \approx \int (X(t) - \mu(t))\phi_k(t)dt,$$

and therefore the conditional expectation (5) approximates the integral definition (2) of the functional principal component scores.

Typical asymptotic results, under appropriate assumptions, include $\lim_{n\to\infty} \widehat{E}[\xi_{ik}|\widetilde{U}_i] = E[\xi_{ik}|\widetilde{U}_i]$, in probability, and for all $t \in \mathcal{T}$, $\lim_{K\to\infty}\lim_{n\to\infty} \widehat{X}_i^K(t) = \widetilde{X}_i(t)$, in probability, where $\widetilde{X}_i(t) = \mu(t) + \sum_{k=1}^{\infty} E[\xi_{ik}|\widetilde{U}_i]\phi_k(t)$ and

$$\widehat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \widehat{E}[A_{ik}|\widetilde{U}_i]\hat{\phi}_k(t) \tag{9}$$

are the predicted individual trajectories, based on K components.

3. Functional regression models and the generalized functional linear model for longitudinal data

One can distinguish various functional regression models according to the nature of predictors and responses. The following scheme lists the most important cases that have been considered.

Predictor	$\mapsto$	Response	Model
$\mathbb{R}^d$		$\mathbb{R}$	Multiple Regression, GLM
$\mathbb{R}^d$		$\mathbb{R}^d$	Multivariate Regression
L^2		L^2	"Functional Regression Model"
$\mathbb{R}^d$		L^2	"Functional Response Model"
L^2		$\mathbb{R}$	"Generalized Functional Linear Model"

The first two of these are classical regression models. In the functional regression model (Ramsay & Dalzell, 1991; Cardot *et al.*, 1999; Cuevas *et al.*, 2002; Cardot *et al.*, 2003) it is assumed that both

predictor and response are random functions. This corresponds to a simple linear regression model for functional data. Extensions of interest would include multiple regression where several functions would serve as predictors, and replacing the linear regression by a nonlinear regression, for example using nearest neighbors or neighborhood metrics in function space to predict a response. Such approaches are discussed in Ferraty & Vieu (2004) and Rice (2004) and open up promising avenues for future research in FDA. A multitude of interesting combinations are conceivable and open to exploration, including additive regression models, multiple index models and sliced inverse regression (Ferre & Yao, 2003). All of these models can be applied to ordinary functional regression, as well as in other areas, such as time series regression (Besse & Cardot, 2002) or survival regression (Müller & Zhang, 2005).

The extension of the simple linear regression model $E(Y|X) = \beta_0 + \beta_1 X$ to functional data $(X(t), Y(t))$ consisting of predictor trajectories X and response trajectories Y is

$$E(Y(t)|X) = \mu(t) + \int X(s)\beta(s,t)\,ds$$

for a "regression parameter function" β. Estimation of the parameter function $\beta(\cdot,\cdot)$ is an inverse problem and requires regularization. Regularization can be implemented in a variety of ways, for example by penalized splines (James, 2002) or by truncation of series expansions (discussed below). A "Functional Normal Equation" (He *et al.*, 2000) extends the multivariate normal equations, and a solution can be obtained under certain conditions. Implementation of this model with sparse longitudinal data has been considered in Yao *et al.*(2005b). A review of functional regression models can be found in Chiou *et al.*(2004), where also a discussion of basic elements of functional diagnostics and goodness-of-fit can be found.

Special cases that merit studies on their own right are the "Functional Response Model" (Chiou *et al.*, 2004), where the response is a random function $Y(t)$ and the predictors Z are scalars or vectors, and the generalized functional linear model (James, 2002; Müller & Stadtmüller, 2005), where the predictors are functions and the responses are generalized variables, of binary, count or continuous type. A simple functional response model that has been termed a functional smooth random effects model (Chiou *et al.*, 2003) is

$$E\{X(t)|Z\} = \mu(t) + \sum_{k=1}^{\infty} E(\xi_k|Z)\phi_k(t),$$

where Z is a possibly multivariate covariate and X is the response function. Further dimension reduction may be achieved by specifying single index models $E(\xi_k|Z) = \alpha_k(\gamma_k' z)$, for $k = 1, 2, \ldots$, where the α_k are smooth link functions, and the γ_k are normalized parameter vectors, in the case of multivariate covariates Z.

A problem closely related to functional regression is functional correlation. To measure the dependency of pairs or vectors of random curves, a natural approach is to extend canonical correlation

(Hotelling, 1936) from MDA to FDA. This involves the inversion of a compact operator (He *et al.*, 2003) and therefore requires sensible regularization, as proposed in Leurgans *et al.*(1993) and further investigated in He *et al.*(2004). The difficulties to find proper regularization prompted a search for alternative functional correlation measures (Service *et al.*, 1998; Heckman & Zamar, 2000; Dubin & Müller, 2005).

Discrimination and classification of curve data is of interest in engineering (Hall and Poskitt, 2001), astronomy (Hall *et al.*, 2000) or for the analysis of gene expression profiles. Discriminant analysis for curve data can be addressed by functional binary regression, which is a special case of generalized functional linear models. Given data $(\{X_i(t)\}, Y_i)$, $i = 1, \ldots, n$, with random predictor curves $X_i(t)$ and a real-valued dependent variable Y, a link function $g(\cdot)$ and a variance function $\sigma^2(\cdot)$, a generalized functional linear model or functional quasi-likelihood model is determined by a parameter function $\beta(\cdot)$, a constant α and linear predictors $\eta = \alpha + \int \beta(t)(X(t) - \mu(t))\, dt$ with conditional means $\mu = g(\eta)$, where $E(Y|X(t)) = \mu$ and $\mathrm{var}(Y|X(t)) = \sigma^2(\mu) = \tilde{\sigma}^2(\eta)$ for a function $\tilde{\sigma}^2(\eta) = \sigma^2(g(\eta))$. The data model is then

$$Y_i = g\left(\alpha + \int \beta(t) X_i(t)\, dt\right) + e_i, \;\; i = 1, \ldots, n, \tag{10}$$

with zero mean errors e_i of appropriate variance structure.

For a given orthonormal basis ϕ_j, $j \geq 1$ of L^2, setting $\zeta = \beta$ in (7) leads to the linear predictor representation $\eta = \alpha + \sum_{j=1}^{\infty} \beta_j \xi_j$ with r.v.'s $\xi_j = \int (X(t) - \mu(t))\phi_j(t)\, dt$ and coefficients $\beta_j = \int \beta(t)\phi_j(t)\, dt$. Regularized estimates are obtained by truncating these expansions at a finite number of terms $K = K(n) \to \infty$ and solving an estimating equation of the type $U(\beta) = 0$. With $\xi^{(i)T} = (\xi_1^{(i)}, \ldots, \xi_p^{(i)})$, $\eta_i = \sum_{j=1}^{K} \beta_j \xi_j^{(i)}$, $\mu_i = g(\eta_i)$, $i = 1, \ldots, n$, the vector valued score function and score equation are defined by

$$U(\beta) = \sum_i (Y_i - \mu_i) g'(\eta_i) \xi^{(i)} / \sigma^2(\mu_i), \qquad U(\beta) = 0. \tag{11}$$

The solutions of this estimating equation then yield the regression function estimates $\hat{\beta}(t) = \hat{\beta}_0 + \sum_j \hat{\beta}_j \phi_j(t)$. Semiparametric extensions with unknown link/variance functions can also be constructed, and under regularity conditions one obtains asymptotic consistency and inference results (Müller & Stadtmüller, 2005).

As an illustration of the PACE principle outlined in section 2, we now may easily extend this model to the case of sparse and noisy longitudinal data as predictors. Namely, since the generalized functional linear model is based on the linear predictor η, a direct application of (8) yields (see (5))

$$\widehat{E}[\eta_i | \widetilde{U}_i] = \alpha + \sum_{k=1}^{K} [\hat{\lambda}_k \hat{\phi}_{ik}^T \hat{\Sigma}_{U_i}^{-1} (\widetilde{U}_i - \hat{\mu}_i)] \beta_k. \tag{12}$$

One applies this method by fitting a regular generalized linear model with the desired variance and link functions by ordinary likelihood or quasi-likelihood with predictors $x_{ki} = \hat{\lambda}_k \hat{\phi}_{ik}^T \hat{\Sigma}_{U_i}^{-1} (\widetilde{U}_i - \hat{\mu}_i)$, yielding estimates for α and β_k, $k = 1, \ldots, K$. The resulting parameter function estimate is then

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{\beta}_k \hat{\phi}_k(t). \tag{13}$$

Any functional regression model that is an extension of a single index or multiple index (Chiou & Müller, 2004) model such as the models proposed by James & Silverman (2004) can be extended in the same way to the case of longitudinal data as predictors.

4. Binomial functional regression for sparse data: An application to the classification of longitudinal data on primary biliary cirrhosis

We consider longitudinal data on patients with primary biliary cirrhosis (PBC), a liver disease. The data resulted from a Mayo Clinic trial that was conducted in 1974 to 1984. They are available at http://lib.stat.cmu.edu/datasets/pbc and have been described in Fleming & Harrington (1991) and Murtaugh *et al.*(1994). The data contain various sparsely and irregularly sampled covariates, as well as the survival or censoring time for each of 312 patients. We consider longitudinal serum bilirubin measurements (in mg/dl), with the aim of predicting long-term survival based on a series of sparse initial measurements. Serum bilirubin concentration is known to be elevated in the presence of chronic liver cirrhosis, such as PBC.

The bilirubin measurements were log-transformed, and only patients were included that survived the first 910 days of the study, our chosen observation period for the bilirubin time courses. Based on the bilirubin time courses during the first 910 days, we then aim at predicting survival beyond 10 years after entering the study. As outcome variable, we introduce a binary indicator variable to indicate whether survival beyond ten years occurred or not. Only patients for whom survival status beyond 10 years was known were retained for this analysis. A total of 258 patients in the study satisfied these criteria. Of these, 84 died between 910 and 3650 days in the study, and 174 lived beyond 10 years. The problem we address is to discriminate between these two groups, solely based on the observed initial time courses of bilirubin. Our approach is to implement a functional binomial regression model using a logistic link, where the binary response corresponds to group membership. This is an example of the functional generalized linear model (10), here implemented for sparse longitudinal data by means of (12), for the special purpose of classifying the observed longitudinal time courses.

The observed initial bilirubin time courses are shown in Fig. 1. The sparse and irregular nature of the bilirubin measurements is evident, as number of measurements, range of the observations and

timing of the measurements differ from patient to patient. These irregularities arise as patients dropped out or were not seen on time for scheduled checks. Longer-living patients generally are found to have lower levels of bilirubin, as expected. No other shape characteristic of the time courses visually stands out. Such plots usually serve as a first descriptive step in FDA, allowing for a visual check of the data, possible identification and removal of outliers, and a preliminary assessment of trends and shapes.

In a second step, these data were subjected to PACE – principal analysis by conditional expectation, as described in section 2, adapting functional PCA to the sparse nature of these measurements. This provides estimates of mean functions, covariance surface, and eigenfunctions. The estimated mean functions are displayed in Fig. 2, separately for the two groups and for the overall mean. We find that the long-lived group has a more or less constant mean bilirubin level over time, with a flat peak around 700d, while the level of the high-risk group is not only higher, but is also increasing over time, with accelerated increases towards 900d.

The smoothed covariance surface in Fig. 3 demonstrates that most of the variation in the trajectories occurs towards the right endpoint, where the covariance function has a peak. The smooth part of the covariance surface is obtained after removing the diagonal of the empirical covariance matrix and a smoothing step. The additional ridge shown along the diagonal corresponds to the effect of measurement error that will be present only along the diagonal. The area between the base and the top of the ridge along the diagonal in Fig. 3, divided by the length where the ridge is visible, provides an estimate for the error variance $\hat{\sigma}^2$ in model (4). If the height of the ridge varies, as it does here, an alternative of interest is to assume a variance function for the error, i.e., heteroscedasticity, rather than a fixed error variance. Bandwidths for covariance smoothing were chosen as $[370d, 370d]$, the smallest bandwidths for which the resulting smoothed covariance surface became positive definite.

Positive definiteness of estimated covariance surfaces is not guaranteed and occasionally can be a problem in practical applications. As in Yao *et al.*(2003), one may then project on positive definite surfaces as follows: Once eigenfunctions and eigenvalues have been determined, one checks whether $\hat{\lambda}_j > 0$ for all estimated eigenvalues. If this is not the case, one simply drops eigenvalue/eigenfunction pairs for which the estimated eigenvalue is negative, and reconstitutes the estimate of the covariance surface from the remaining eigenvalue/eigenfunction estimates. Then the modified covariance surface estimate is

$$\widetilde{C}(s,t) = \sum_{j:\hat{\lambda}_j>0}^{K} \hat{\lambda}_j \hat{\phi}_j(s) \hat{\phi}_j(t).$$

The first three eigenfunctions resulting from PACE, based on covariance and mean function estimates, are displayed in Fig. 4. In accordance with the behaviour of the covariance surface, most of the variation of the eigenfunctions is concentrated towards the right end. The estimated eigenfunctions are

approximately orthonormal, which is reflected in their increasing bumpiness as their order increases; indeed, it is easy to show that additional sign changes are needed in order to maintain orthogonality as the order increases. These eigenfunctions correspond to the "modes of variation" of the data and indicate in which functional directions the sample of functions will vary the most. First and second eigenfunctions are fairly flat in the beginning and differentiate time courses only in the right half of the domain, where the first eigenfunction picks up rapidly rising time courses, while the second eigenfunction is aligned with declining time courses (or vice versa, as the sign is not uniquely determined).

One-curve-leave-out cross-validation aiming at minimizing prediction error has been suggested for the choice of the number of components K (Rice & Silverman, 1991). Define $\hat{\mu}^{(-i)}$, $\hat{\phi}_k^{(-i)}$ and $\widehat{Y}_i^{(-i)}$ as estimated mean function, eigenfunctions, and predicted trajectory for the i-th subject, obtained by omitting the data for the ith subject. The cross-validation choice $\hat{K}$ then is

$$\hat{K} = \text{argmin}_K \sum_{i=1}^{n} \sum_{j=1}^{N_i} \{U_{ij} - \widehat{X}_i^{(-i)}(T_{ij})\}^2.$$

Computationally faster alternatives can be obtained via a pseudo-Gaussian likelihood (Yao *et al.*, 2005a), conditional on the predicted functional principal component scores $\tilde{\xi}_{ik}$. Using the same notation as in (5), this pseudo-Gaussian likelihood is given by

$$\widehat{L} = \sum_{i=1}^{n} \{-\frac{N_i}{2} \log(2\pi) - \frac{N_i}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} (\widetilde{U}_i - \hat{\mu}_i - \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_{ik})^T (\widetilde{U}_i - \hat{\mu}_i - \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_{ik})\}, \qquad (14)$$

where minimization of AIC $= -\widehat{L} + K$ and BIC $= -\widehat{L} + K \log n$ then leads to implementations of AIC and BIC model selectors. All of these criteria, when applied to the data of our example, point to two as a reasonable number of eigenfunctions, and this choice was implemented in the subsequent analysis.

Predicted log(bilirubin) trajectories for eight patients with sparse measurements can be viewed in Fig. 5. Predicted functional principal component scores $\tilde{\xi}_{i1}, \tilde{\xi}_{i2}$ (6) are used to construct predicted trajectories $\widehat{X}(t) = \tilde{\xi}_{i1}\hat{\phi}_1(t) + \tilde{\xi}_{i2}\hat{\phi}_2(t)$. Overall, the obtained fits appear reasonable in the interior, while some of the bumpier fits near the right boundary are less well supported by the observations themselves and seem to involve some extrapolation.

The estimated probabilities of belonging to the long-lived group are also provided in Fig. 5. These are obtained in a last step by solving the score equation (11) which can be done easily with iterated weighted least squares or with generally available software, by running a binomial regression with predictors $\tilde{\xi}_{ik}$, where we choose here the logit link. We pair patients with similar shapes of predicted trajectories in adjacent left and right panels and note that the survival probability is mainly determined by the log(bilirubin) level, with a declining trend being slightly advantageous.

The predicted functional principal component scores $\tilde{\xi}_{i1}, \tilde{\xi}_{i2}$ themselves are illustrated in Fig. 6. This figure indicates that the discrimination task is not straightforward for these data; there exists substantial overlap between the two groups. Using two eigenfunctions, Table 1 provides the sample statistics for the coefficient estimates $\hat{\beta}_k$, $k = 0, 1, 2$, that are obtained from the logistic regression of group indicator (long-lived =1, short-lived =0) on the predicted functional principal component scores $\tilde{\xi}_k$, $k = 1, 2$, with linear predictor $\eta = \alpha + \beta_1\tilde{\xi}_1 + \beta_2\tilde{\xi}_2$.

Table 1: Parameter estimates from logistic model fit for classifying PBC data.

	α	β_1	β_2
Estimate	0.8583	-0.0256	-0.0293
std.err.	0.0226	<0.0001	0.0001
p-value	<0.0001	0.0003	0.0171

Table 2: One-leave out classification results for functional logistic model in PBC study.

	True	
Classified	Long Survival	Short Survival
Long	126	28
Short	50	56

Putting everything together, we can then predict group membership from the observed sparse data. Using the one-leave-out prediction error criterion, the overall misclassification rate for this procedure is 26.54%. Further details are in Table 2.

5. Outlook

Functional data analysis is a rapidly evolving field. There are many problems left to be addressed in future research. Theoretical developments supporting FDA methodology are still in an initial stage and there exists a need to develop realistic functional asymptotics, especially inference. Using flexible parametric models such as B-splines with finitely many components is a parametric short-cut that allows to obtain inference easily due to its fully parametric nature and may lead to satisfactory results in various applications. Such approaches however do not reflect the functional nature of the problem and are not providing the correct asymptotics, as long as the number of knots is kept fixed.

Other open problems concern the application of FDA methods to samples of functions that are not

traditional random trajectories. Larger classes of objects than have been previously addressed with these methods, which traditionally have been densely sampled random trajectories observed without error, are targeted by new versions of FDA. Extensions to irregularly measured and noisy trajectories have been discussed in this paper. Examples of objects with special properties that are of interest are samples of density functions (Kneip & Utikal, 2001) and hazard functions (Müller *et al.*, 1997). Non-continuous measurements such as binomial or Poisson distributed repeated measurements provide another class of interest. Objects such as random surfaces and higher-dimensional functions in image analysis and also the analysis of shapes, texts, gene sequences and textures are possible targets of suitably modified FDA techniques.

Application areas of FDA are likely to encompass many fields of statistics where scalar or multivariate observations are complemented by observations in the form of random trajectories, such as time-course gene expression data (Aach & Church, 2001; Liu & Müller, 2003; Zhao *et al.*, 2004), physiological time courses (Ratcliffe *et al.*, 2002) or industrial and quality control problems (Faraway, 1997; Woodall *et al.*, 2004).

Acknowledgements

This research was supported in part by NSF grants DMS02-04869 and DMS03-54448.

References

Aach, J. & Church, G.M. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics* **17**, 495-508.

Besse, P.C. & Cardot, H. (2002). Spline approximation of the forecast of a first-order autoregressive functional process. *Canad. J. Statist.* **24**, 467-487.

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

Bosq, D. (1991). Modelization, nonparametric estimation and prediction for continuous time processes. In *Nonparametric functional estimation and related topics* (1991 ed.), ed. G. Roussas, Dordrecht, Netherlands: Kluwer Academic, pp 509-529.

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

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparam. Statist.* **12**, 503-538.

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. (2004). Quasi-likelihood regression with multiple indices and smooth link and variance functions. *Scand. J. Statist.* **31**, 367-386.

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.

Courant, R. & Hilbert, D. (1953). *Methods of mathematical physics* (1989 ed.), New York: Wiley.

Cuevas, A., Febrero, M. & Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *Can. J. Statist.* **30**, 285-300.

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.

Davidian, M. & Giltinan, D.M. (1995). Nonlinear models for repeated measurement data. Chapman & Hall, London

Dubin, J. & Müller, H.G. (2005). Dynamical correlation for multivariate longitudinal data. *J. Amer. Statist. Assoc.*, to appear.

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. (2004). Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. *J. Nonpara. Statist.* **16**, 111-125.

Ferre, L. & Yao, A.F. (2003). Functional sliced inverse regression analysis. *Statistics* **37**, 475-488.

Fleming, T.R. & Harrington, D.P. (1991). *Counting processes and survival analysis.* Wiley, 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., 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.

Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv för Matematik*, 195-276.

Hall, P. & Poskitt, D.S. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43**, 1-9.

Hall, P., Reimann, J. & Rice, J. (2000). Nonparametric estimation of a periodic function. *Biometrika* **87**, 545-557.

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.

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

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educational Psychology* **24**, 417-441, 498-520.

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.

James, G. & Silverman, B.W. (2004). Functional adaptive model estimation. Preprint.

Joliffe, I.T. (2002). *Principal component analysis.* Springer, New York.

Jones, R.H. (1993). Longitudinal data with serial correlation. Chapman & Hall, London.

Karhunen, K. (1946). Zur Spektraltheorie stochastischer Prozesse. *Ann. Acad. Sci. Fennicae* **A I 37**.

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.

Kneip, A., Li, X., MacGibbon, K.B. & Ramsay, J.O. (2000). Curve registration by local regression. *Can. J. Statist.*, **28**, 19-29.

Kneip, A. & Utikal, K.J. (2001). Inference for density families using functional principal component analysis analysis. *J. Amer. Statist. Assoc.* **96**, 519-532.

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.

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.

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.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis*, London: Academic Press.

Müller, H.G., Wang, J.L., Capra, W.B., Liedo, P., Carey, J.R. (1997). Early mortality surge in protein-deprived females causes reversal of sex differential of life expectancy in Mediterranean fruit flies. *Proceedings of the National Academy of Sciences USA* **94**, 2762-2765.

Müller, H.G. & Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.*, to appear.

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

Murtaugh, P.A., Dickson, E.R., Van Dam, G.M., Malinchoc, M., Grambsch, P.M., Langworthy, A.L. & Gips, C.H. (1994). Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology* **20**, 126-136.

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. (1997). *Functional data analysis*, New York: Springer.

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

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

Ratcliffe, S.J., Leader, L.R. & Heller, G.Z. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression. *Statist. Med.* **21**, 1103-1114.

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.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric regression.* Cambridge University Press

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.

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

Stützle, W., Gasser, T., Molinari, L., Largo, R., Prader, A. & Huber, P.J. (1980). Shape-invariant modeling of human growth. *Ann. Human Biology* **7**, 507-528.

Wang, K.M. & Gasser, T. (1997). Alignment of curves by dynamic time warping. *Ann. Statist.*, **25**, 1251-1276.

Wang, K.M. & Gasser, T. (1998). Asymptotic and bootstrap confidence bounds for the structural average of curves. *Ann. Statist.*, **26**, 972-991.

Wang, K.M. & Gasser, T. (1999). Synchronizing sample curves nonparametrically. *Ann. Statist.*, **27**, 439-460.

Woodall, W.H., Spitzner, D.J., Montgomery, D.C. & Gupta, S. (2004). Using control charts to monitor process and product quality profiles. *J. Quality Technology* **36**, 309-320.

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. Amer. Statist. Assoc.*, to appear.

Yao, F., Müller, H.G., Wang, J.L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.*, to appear.

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

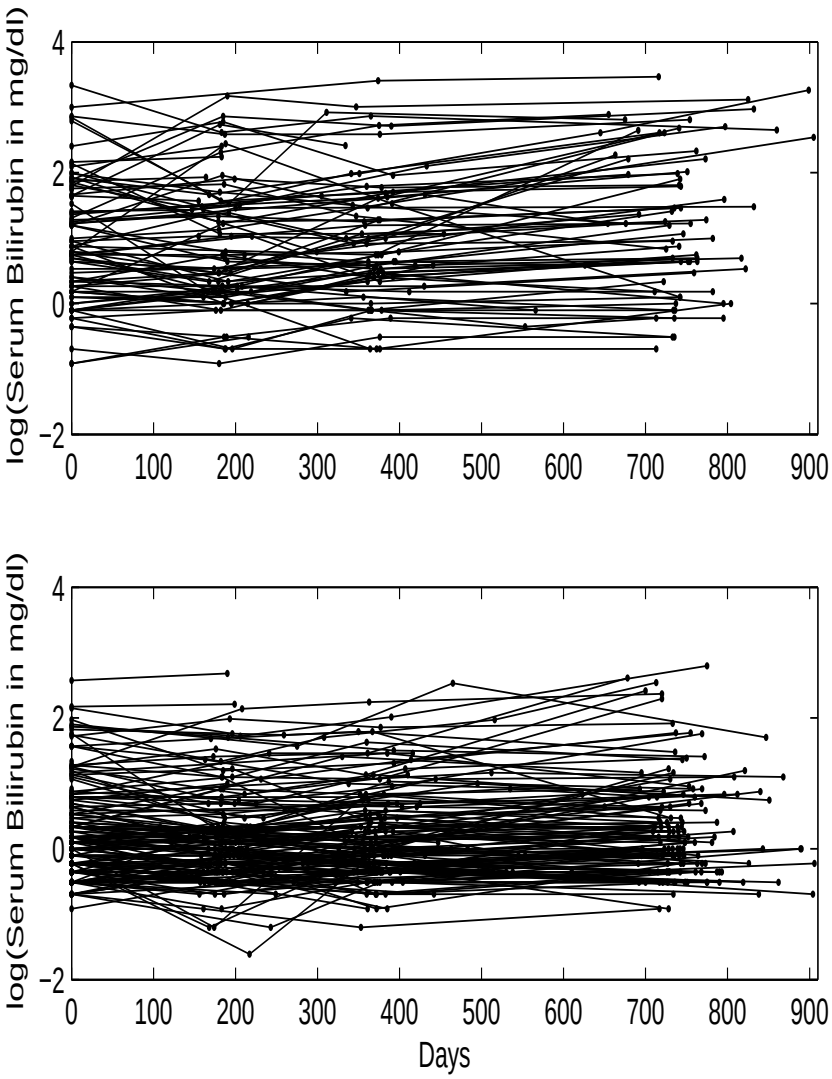

Figure 1: The observed time courses of log(serum bilirubin) during the first 910 days in the study, for 84 short-lived (upper panel) and 174 long-lived (lower panel) patients from the PBC study.

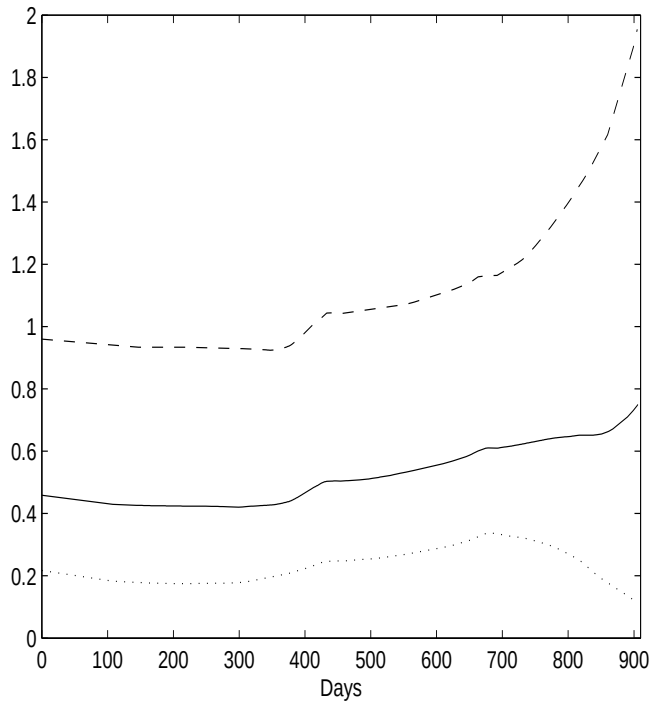

Figure 2: Smoothed mean functions for all patients in the sample (solid), the long-lived patients (dotted) and the short-lived patients (dashed) (with bandwidth 300d). The values on the y axis are log-transformed bilirubin levels.

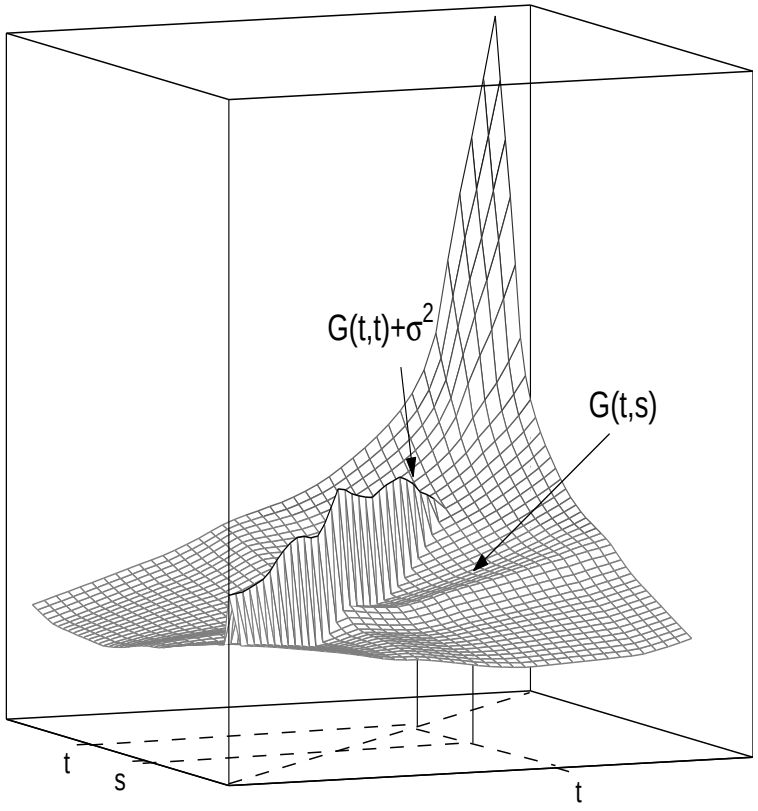

Figure 3: Smoothed covariance function $G(s,t)$, obtained by omitting the diagonal terms (with bandwidths [370d,370d]). Overlaid is the discontinuous ridge along the diagonal that is caused by additional noise, where the tip of the ridge corresponds to $G(t,t)+\sigma^2$ and the base to $G(t,t)$.

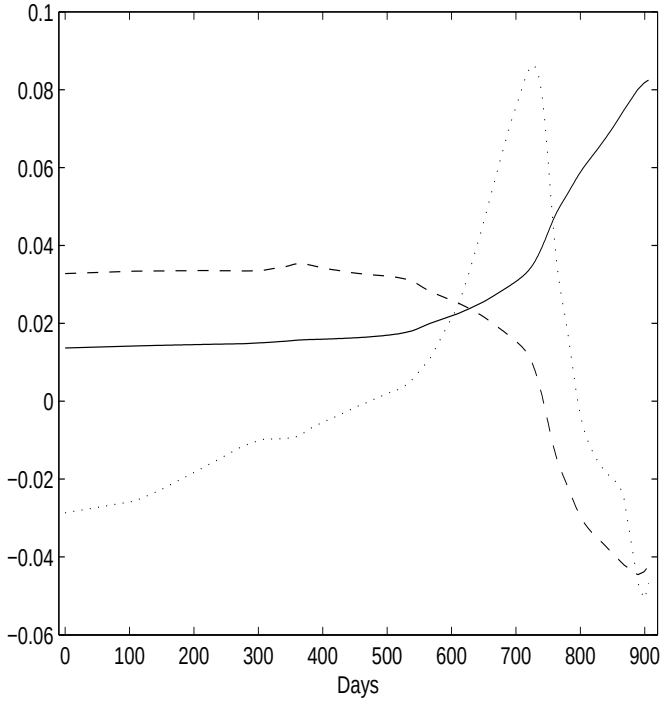

Figure 4: First (solid), second (dashed) and third (dotted) smoothed eigenfunctions for the PBC data, explaining, respectively, 79.3%, 17.6% and 2.1% of total variance.

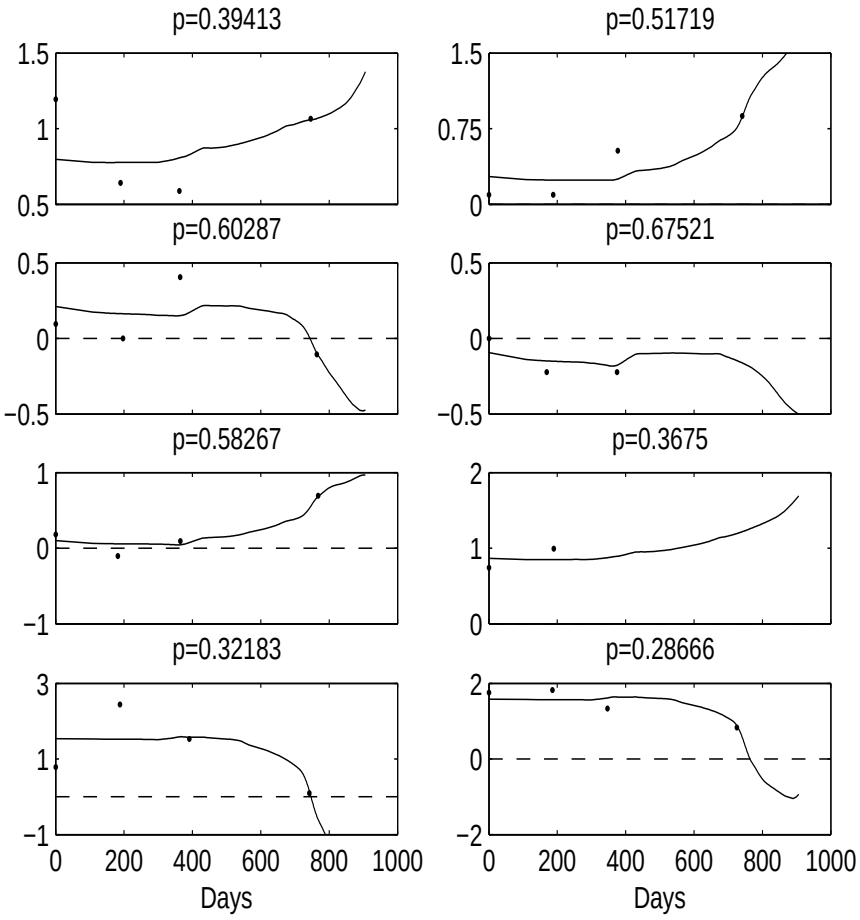

Figure 5: Observed sparse data, predicted log(bilirubin) trajectories and predicted probabilities for belonging to the long-lived group for eight randomly selected patients of the PBC study.

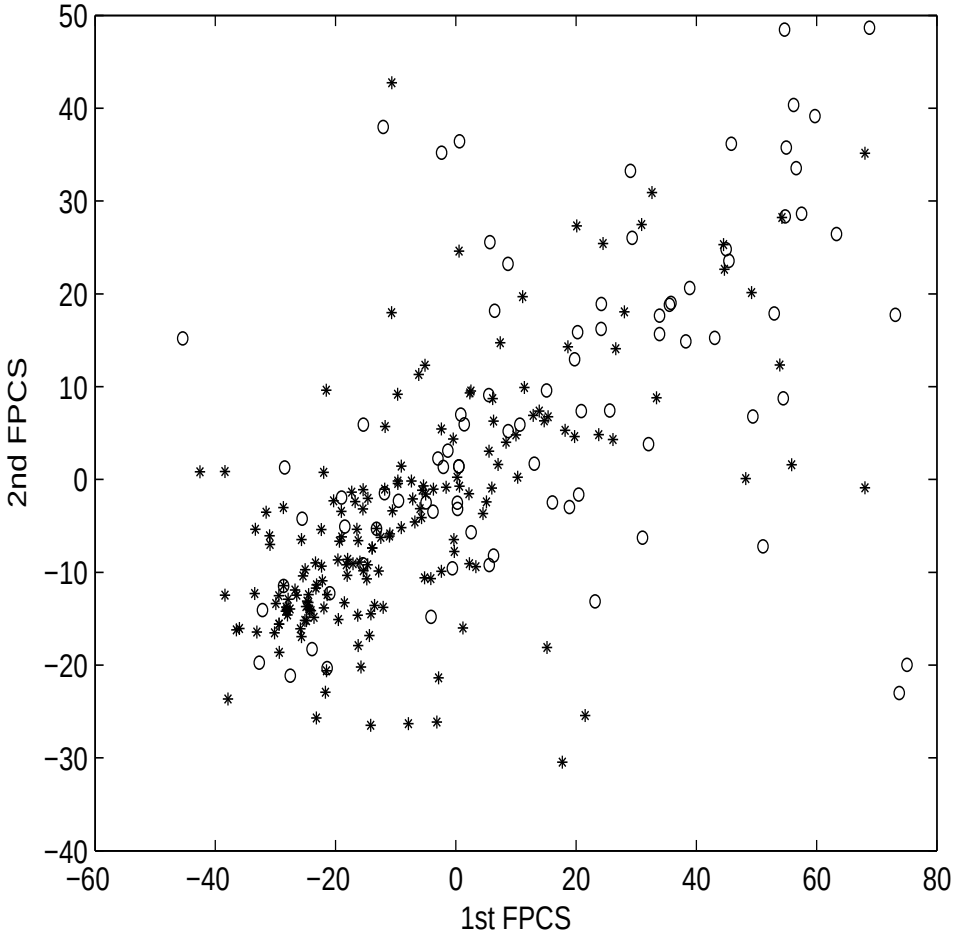

Figure 6: Second versus first predicted functional principal component scores for PBC data. The crosses indicate long-lived and the circles short-lived patients.