

# Sparseness and Functional Data Analysis

Gareth JAMES  
Marshall School of Business  
University of Southern California, Los Angeles, California  
gareth@usc.edu

## Abstract

In this chapter we examine two different settings in which sparseness can be important in a functional data analysis (FDA). The first setting involves sparseness in the functions. The classical assumption of FDA is that each function has been measured at all time points. However, in practice it is often the case that the functions have only been observed at a relatively small number of points. Here we discuss different general approaches that can be applied in this setting, such as basis functions, mixed effects models and local smoothing, and examine the relative merits of each. Then we briefly outline several specific methodologies that have been developed for dealing with sparse functional data in the principal components, clustering, classification and regression paradigms. The second setting involves using sparsity ideas from high dimensional regression problems, where most of the regression coefficients are assumed to be zero, to perform a dimension reduction in the functional space. We discuss two specific approaches that have been suggested in the literature.

*Key Words:* Basis functions; Dantzig Selector; FLiRTI; Functional classification; Functional clustering; Functional principal components; Functional regression; High dimensional regression; Lasso; Local smoothing; Mixed effects models; PACE; Regularization; SASDA; Smoothing spline; Sparse functional data.

## 1 Introduction

It is often assumed in functional data analysis (FDA) that the curve or function has been measured over all points or, more realistically, over a densely sampled grid. In this chapter we deal with two FDA settings where sparsity becomes important. We first consider the situation where, rather than densely observed functions, we only have measurements at a relatively

sparse set of points. In this situation alternative methods of analysis are required.

Figure 1 provides an example of two data sets that fall into the “sparsely observed” category. The “growth” data, illustrated in Figure 1a), consists of measurements of spinal bone mineral density for 280 males and females taken at various ages and is a subset of the data presented in Bachrach *et al.* (1999). Even though, in aggregate, there are 860 observations taken over a period of almost two decades, there are only 2-4 measurements for each individual covering no more than a few years. The “Nephropathy” data, illustrated in Figure 1b), shows percentage changes in glomular filtration rate (GFR) over a six year period, for a group of patients with membranous nephropathy, an auto-immune disease of the kidney. GFR is a standard measure of kidney function. Both of these data sets have sparsely observed curves so more standard FDA techniques can not be applied. We will use both data sets to illustrate alternative approaches that have been developed for dealing with sparse functional data. In Section 2 we discuss different classes of methods that can be applied to functional data and examine the relative merits of each approach for sparsely observed data. Then in Section 3 we briefly outline several specific methodologies that have been developed for dealing with sparse functional data in the principal components, clustering, classification and regression settings.

Section 4 considers the second situation where sparsity plays a key role. Here we examine two approaches that utilize variable selection ideas from the high dimensional regression literature to perform functional regressions. The first approach uses variable selection methods to identify a small set of time points that are jointly most predictive for the response,  $Y$ . A local linear estimator is then used to form predictions for  $Y$  based on the observed values of the predictor,  $X(t)$ , at the selected time points. The second approach performs an interpretable functional linear regression by assuming sparsity in the derivatives of  $\beta(t)$ . For example, a simple piecewise linear estimate for  $\beta(t)$  can be produced provided  $\beta''(t)$  is zero at “most” time points. Variable selection methods are then used to identify the locations where  $\beta''(t)$  is non-zero.

## 2 Different Approaches To Functional Data

In the classic FDA situation with densely observed curves a common approach involves forming a fine grid of time points and sampling the curves at each time point. This approach results in a high, but finite, dimensional

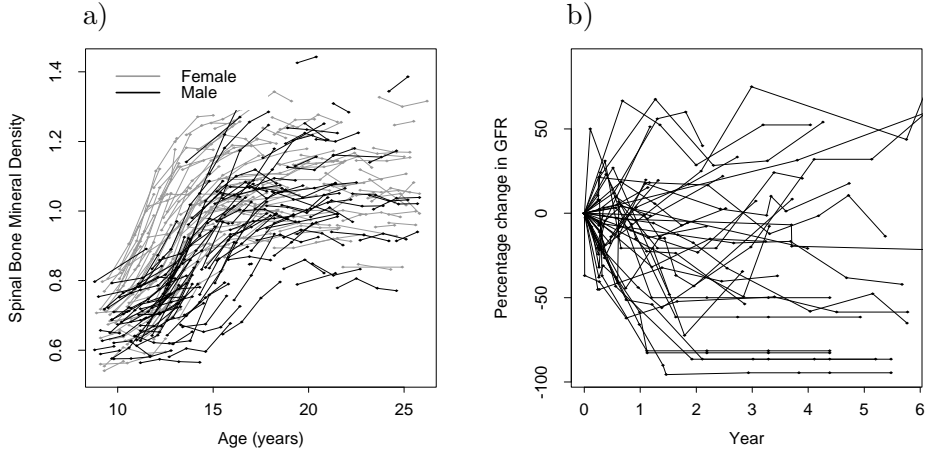


Figure 1: a) Measurements of spinal bone mineral density ( $g/cm^2$ ) for males (black) and females (grey) at various ages,  $n = 280$ . b) Percentage change in GFR scores for 39 patients with membranous nephropathy.

representation for each curve. One can then apply standard statistical methods to the resulting data vectors. Since the vectors are high dimensional and nearby observations in time will generally be highly correlated the resulting covariance matrices can be unstable. A common solution involves imposing some kind of regularization constraint (DiPillo, 1976; Friedman, 1989; Banfield and Raftery, 1993; Hastie *et al.*, 1995). While this discretization followed by regularization approach can work well on densely observed functional data it will fail for the types of sparse data illustrated in Figure 1. For this data, each individual has been observed at different time points. Hence, any fixed grid that is formed will involve many missing observations for each curve. For such data a new approach becomes necessary. In Sections 2.1, 2.2 and 2.3 we discuss three possible alternatives. Throughout this section we assume that one observes  $N$  functions,  $Y_1(t), \dots, Y_N(t)$  with the  $i$ th function observed at time points  $t_{i1}, \dots, t_{in_i}$  and  $Y_{ij} = Y_i(t_{ij})$ .

## 2.1 Basis Functions

A natural way to model the smooth shape of each curve observed over time is to choose a finite  $p$ -dimensional basis,  $s_\ell(t)$ ,  $\ell = 1, \dots, p$ . One can then

assume

$$Y_i(t) = \sum_{l=1}^p s_l(t)\eta_{il} = \mathbf{s}(t)^T \boldsymbol{\eta}_i \quad (1)$$

where  $\mathbf{s}(t) = [s_1(t), s_2(t), \dots, s_p(t)]^T$  and  $\boldsymbol{\eta}_i$  represents the basis coefficients for the  $i$ th curve. Common examples for  $s_\ell(t)$  include spline and Fourier bases. Once the basis and dimension,  $p$ , have been chosen the  $\boldsymbol{\eta}_i$ 's can be estimated by applying standard least squares, separately to each curve, using the linear model (1).

Once the  $\boldsymbol{\eta}_i$ 's have been estimated there are two common approaches to analyzing the data. The first involves using the estimated  $\boldsymbol{\eta}_i$  and (1) to provide an estimate of the entire curve,  $Y_i(t)$ . One can then treat the estimated curve as if it had been observed and apply the discretization approach previously discussed. Alternatively, one can treat the  $p$ -dimensional estimated  $\boldsymbol{\eta}_i$ 's as the observed data and apply standard statistical methods. A related approach to the basis method involves fitting a separate smoothing spline to each curve by finding  $g_i(t)$  to minimize

$$\sum_{j=1}^{n_i} (Y_i(t_{ij}) - g_i(t_{ij}))^2 + \lambda \int (g_i''(t))^2 dt.$$

Once the  $g_i$ 's have been estimated the discretization approach can be applied.

The basis and smoothing methods both have the advantage over the straight discretization approach that they can be applied to data where the curves have not all been measured at the same time points. For example, they can be used on the Nephropathy data. However, the basis method also has several problems. First, when the curves are measured at different time points, it is easy to show that the variance of the estimated basis coefficients,  $\hat{\boldsymbol{\eta}}_i$ , is different for each individual. Hence, any statistical analysis that utilizes the  $\hat{\boldsymbol{\eta}}_i$ 's should place more weight on the more accurately estimated basis coefficients. However, it is often not obvious how one should treat the  $\hat{\boldsymbol{\eta}}_i$ 's differently. Even more importantly, for extremely sparse data sets many of the basis coefficients will have infinite variance, making it impossible to produce reasonable estimates. For example, in the growth data there are so few observations that it is not possible to fit a separate curve for each individual using any reasonable basis. In this case the basis approach will fail.

## 2.2 Mixed Effects Methods

The main reason that the basis approach fails for the growth data is that it only uses the information from a particular curve to estimate the basis coefficients for that curve. For the growth data there are few observations per curve but a large number of curves so in total we still have a lot of information. Hence, a potentially superior method would be to somehow utilize the information from all curves to estimate the coefficients for each curve. A natural way to achieve this goal is to utilize a mixed effects framework. Mixed effects models have been widely used in the analysis of functional data; see for instance Shi *et al.* (1996), Brumback and Rice (1998) and Rice and Wu (2001) for some early examples.

Denote by  $\boldsymbol{\beta}$  an unknown but fixed vector of spline coefficients, let  $\boldsymbol{\gamma}_i$  be a random vector of spline coefficients for each curve with population covariance matrix  $\boldsymbol{\Gamma}$ , and let  $\epsilon_i(t)$  be random noise with mean zero and variance  $\sigma^2$ . The resulting mixed effects model has the form

$$Y_i(t) = \mathbf{s}(t)^T \boldsymbol{\beta} + \mathbf{s}(t)^T \boldsymbol{\gamma}_i + \epsilon_i(t) \quad i = 1, \dots, N. \quad (2)$$

In practice  $Y_i(t)$  is only observed at a finite set of time points. Let  $\mathbf{Y}_i$  be the vector consisting of the  $n_i$  observed values, let  $\mathbf{S}_i$  be the corresponding  $n_i$  by  $p$  spline basis matrix evaluated at these time points, and let  $\boldsymbol{\epsilon}_i$  be the corresponding random noise vector with covariance matrix  $\sigma^2 \mathbf{I}$ . The mixed effects model then becomes

$$\mathbf{Y}_i = \mathbf{S}_i \boldsymbol{\beta} + \mathbf{S}_i \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i \quad i = 1, \dots, N. \quad (3)$$

The fixed-effects term,  $\mathbf{S}_i \boldsymbol{\beta}$ , models the mean curve for the population and the random-effects term,  $\mathbf{S}_i \boldsymbol{\gamma}_i$ , allows for individual variation. The principal patterns of variation about the mean curve are referred to as functional principal component curves.

A general approach to fitting mixed effects models of this form uses the EM algorithm to estimate  $\boldsymbol{\beta}$ ,  $\boldsymbol{\Gamma}$  and  $\sigma^2$  (Laird and Ware, 1982). Given these estimates, predictions are obtained for the  $\boldsymbol{\gamma}_i$ 's using the "best linear unbiased prediction" (Henderson, 1950). For (3) above, the best linear unbiased prediction for  $\boldsymbol{\gamma}_i$  is

$$\hat{\boldsymbol{\gamma}}_i = (\hat{\boldsymbol{\Gamma}}^{-1} / \hat{\sigma}^2 + \mathbf{S}_i^T \mathbf{S}_i)^{-1} \mathbf{S}_i^T (\mathbf{Y}_i - \mathbf{S}_i \hat{\boldsymbol{\beta}}).$$

Once the  $\hat{\boldsymbol{\gamma}}_i$ 's have been computed one can then either form predictions for each  $Y_i(t)$  using (2) and utilize the discretization approach, or else simply treat the  $\hat{\boldsymbol{\gamma}}_i$ 's as the observed  $p$ -dimensional data.

The mixed effects method has many advantages over the basis approach. First, it estimates the curve  $Y_i(t)$  using all the observed data points rather than just those from the  $i$ th individual. This means that the mixed effects method can be applied when there are insufficient data from each individual curve to use the basis method. For example, James *et al.* (2000) successfully use a mixed effects model to fit the growth data. Secondly, it uses maximum likelihood to estimate the parameters. Thus it automatically assigns the correct weight to each observation and the resulting estimators have all the usual asymptotic optimality properties.

### 2.3 Local Smoothing

The mixed effects approach is not the only method for building strength across functions by incorporating information from all the curves. An alternative approach, that does not require specifying basis functions, involves using local smoothing techniques to estimate the mean and covariance functions of the curves. The smoothing approach makes use of the Karhunen-Loeve expansion which states that any smooth curve can be decomposed as  $Y(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$  where  $\xi_k$  is the  $k$ th principal component score and  $\phi_k(t)$  is the  $k$ th principal component function. Hence a natural approximation for  $Y(t)$  is given by

$$\hat{Y}_i(t) = \hat{\mu}(t) + \sum_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t) \quad (4)$$

where  $K$  represents the number of principal components used in the estimation. Lower values of  $K$  correspond to more regularization. Next we outline the method for estimating  $\hat{\mu}(t)$ ,  $\hat{\xi}_k$  and  $\hat{\phi}_k(t)$ .

Let  $Y_i(t) = \mu(t) + X_i(t) + \epsilon_i(t)$  where  $\mu(t) = EY_i(t)$  and  $Var(\epsilon_i) = \sigma^2$ . Further, let  $G(s, t) = cov(X(s), X(t))$ , the covariance function of  $X(t)$ . To implement this approach one first pools all the observations for  $Y_1(t), \dots, Y_N(t)$  and uses a kernel method such as a local linear smoother to estimate  $\mu(t)$ . Next, the estimated mean is subtracted from the curves and the “raw” covariance estimates

$$\hat{G}_i(t_{ij}, t_{il}) = (Y_{ij} - \hat{\mu}(t_{ij}))(Y_{il} - \hat{\mu}(t_{il}))$$

are computed. A local smoother is then applied to all the  $\hat{G}_i(t_{ij}, t_{il})$  points where  $j \neq l$  to produce,  $\hat{G}(s, t)$ , an estimate for  $G(s, t)$ . From the estimated covariance function one can compute the corresponding eigenvalues,  $\hat{\lambda}_k$  and

eigenfunctions,  $\hat{\phi}_k(t)$  via

$$\int \widehat{G}(s, t) \hat{\phi}_k(s) ds = \hat{\lambda}_k \hat{\phi}_k(t) \quad (5)$$

where  $\int \hat{\phi}_k(t)^2 dt = 1$  and  $\int \hat{\phi}_k(t) \hat{\phi}_m(t) dt = 0$  for  $m < k$ . Equation (5) can be estimated by evaluating  $\widehat{G}(s, t)$  over a fine grid and then computing the standard eigenvectors and eigenvalues.

The final step in implementing (4) requires the estimation of  $\hat{\xi}_{ik}$ . Yao *et al.* (2005a) recommend using an approach they call PACE which involves computing the conditional expectation  $E[\xi_{ik} | \mathbf{Y}_i]$  where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ . They show that

$$E[\xi_{ik} | \mathbf{Y}_i] = \lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \quad (6)$$

where  $\boldsymbol{\phi}_{ik}$  and  $\boldsymbol{\mu}_i$  are vectors corresponding to the principal component function and mean function evaluated at  $t_{i1}, \dots, t_{in_i}$  and  $(\boldsymbol{\Sigma}_i)_{jl} = G(t_{ij}, t_{il}) + \sigma^2 \delta_{jl}$  where  $\delta_{jl} = 1$  if  $j = l$  and 0 otherwise. Plugging in the estimates of these various parameters gives  $\hat{\xi}_{ik}$ . Once the  $\hat{\xi}_{ik}$ 's have been estimated one can either compute the  $Y_i(t)$ 's using (4) and apply the discretization approach or else treat the  $\hat{\xi}_{ik}$ 's as a  $K$ -dimensional representation of the curves. See Yao *et al.* (2005a) for an application of this approach to estimate functional principal components.

Both the basis function and local smoothing approaches require the choice of tuning parameters. For basis functions one must select the dimension of the basis and for local smoothing the bandwidth. As with all problems involving tuning parameters there are a number of possible approaches that can be used such as AIC, BIC or cross-validation. In practice, depending on the sparsity level of the data, visual inspection is often used to select a reasonable tradeoff between fit to the data and smoothness of the curves.

### 3 Applications of FDA to Sparse Data

This section illustrates some applications of the functional data analysis paradigm to sparse data situations. In particular we discuss functional principal components, functional clustering, functional classification and functional regression methods.

#### 3.1 Functional Principal Components Analysis

A number of papers have been written on the problem of computing functional principal components (FPCA). Ramsay and Silverman (2005) discuss

FPCA for densely sampled curves (see also Chapter 8). For sparsely sampled data a few references include Shi *et al.* (1996), Staniswalis and Lee (1998), James *et al.* (2000), Rice and Wu (2001), and Yao *et al.* (2005a). These approaches tend to either use the mixed effects framework of Section 2.2 or the local smoother method from Section 2.3. For example the approach of Yao *et al.* (2005a) first uses a local smoother to estimate the covariance function,  $G(s, t)$ . The covariance function is then discretized over a fine grid from which the eigenvectors and eigenvalues are computed. Finally, the principal component scores,  $\xi_{ik}$ , are computed using the PACE estimate given by (6).

Here we outline an alternative approach to that of Yao *et al.* (2005a) which utilizes the mixed effects framework discussed in Section 2.2. The mixed effects approach to functional principal components has been used in several papers (Shi *et al.*, 1996; James *et al.*, 2000; Rice and Wu, 2001). Let  $Y_i(t)$  be the measurement at time  $t$  for the  $i$ th individual or curve. Let  $\mu(t)$  be the overall mean function, let  $f_j$  be the  $j$ th principal component function and set  $f = (f_1, f_2, \dots, f_K)^T$ . To estimate  $K$  principal component curves we first define a general additive model

$$\begin{aligned} Y_i(t) &= \mu(t) + \sum_{j=1}^K f_j(t)\alpha_{ij} + \epsilon_i(t) \\ &= \mu(t) + f(t)^T \boldsymbol{\alpha}_i + \epsilon_i(t) \quad i = 1, \dots, N, \end{aligned}$$

subject to the orthogonality constraint  $\int f_j f_l = 0$  for  $j \neq l$  and 1 otherwise. The random vector  $\boldsymbol{\alpha}_i$  gives the relative weights on the principal component functions for the  $i$ th individual and  $\epsilon_i(t)$  is random measurement error. The  $\boldsymbol{\alpha}_i$ 's and  $\epsilon_i$ 's are all assumed to have mean zero. The  $\boldsymbol{\alpha}_i$ 's are taken to have a common covariance matrix,  $\boldsymbol{\Sigma}$ , and the measurement errors are assumed uncorrelated with a constant variance of  $\sigma^2$ .

In order to fit this model when the data are measured at only a finite number of time points it is necessary to place some restrictions on the form of the mean and principal component curves so one represents  $\mu$  and  $f$  using a finite dimensional basis, such as spline functions (Silverman, 1985; Green and Silverman, 1994). Let  $\mathbf{s}(t)$  be a  $p$ -dimensional orthogonal basis. Let  $\boldsymbol{\Theta}$  and  $\boldsymbol{\theta}_\mu$  be, respectively, a  $p$  by  $k$  matrix and a  $p$ -dimensional vector of spline coefficients. Then  $\mu(t) = \mathbf{s}(t)^T \boldsymbol{\theta}_\mu$ , and  $f(t)^T = \mathbf{s}(t)^T \boldsymbol{\Theta}$ . The resulting restricted model has the form

$$\begin{aligned} Y_i(t) &= \mathbf{s}(t)^T \boldsymbol{\theta}_\mu + \mathbf{s}(t)^T \boldsymbol{\Theta} \boldsymbol{\alpha}_i + \epsilon_i(t), \quad i = 1, \dots, N, \\ \epsilon_i(t) &\sim N(0, \sigma^2), \quad \boldsymbol{\alpha}_i \sim N(\mathbf{0}, \mathbf{D}) \end{aligned}$$

subject to

$$\Theta^T \Theta = \mathbf{I}, \quad \int \mathbf{s}(t)^T \mathbf{s}(t) dt = 1, \quad \int \int \mathbf{s}(t)^T \mathbf{s}(s) dt ds = 0. \quad (7)$$

The equations in (7) impose orthogonality constraints on the principal component curves. Note that, if one does not assume a special structure for the covariance matrix of the  $\alpha_i$ 's,  $\Theta$  and  $\Sigma$  are confounded. Thus we restrict the covariance matrix to be diagonal and denote it by  $\mathbf{D}$ .

For each individual  $i$ , let  $t_{i1}, t_{i2}, \dots, t_{in_i}$  be the time points at which measurements are available. Then  $\mathbf{Y}_i = (Y_i(t_{i1}), \dots, Y_i(t_{in_i}))^T$ , and  $\mathbf{S}_i = (\mathbf{s}(t_{i1}), \dots, \mathbf{s}(t_{in_i}))^T$ . Note that  $\mathbf{S}_i$  is the basis matrix for the  $i$ th individual. To approximate the orthogonality condition in (7) we choose  $\mathbf{s}(\cdot)$  so that  $\mathbf{S}^T \mathbf{S} = \mathbf{I}$ , where  $\mathbf{S}$  is the basis matrix evaluated on a fine grid of time points. For instance, for the growth data the time interval was divided into 172 periods of 1/10th of a year each.

The final model can then be written as

$$\mathbf{Y}_i = \mathbf{S}_i \boldsymbol{\theta}_\mu + \mathbf{S}_i \Theta \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, N, \quad (8)$$

$$\Theta^T \Theta = \mathbf{I}, \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad \boldsymbol{\alpha}_i \sim N(\mathbf{0}, \mathbf{D}).$$

James *et al.* (2000) provide an EM fitting procedure for this model and also suggest methods for choosing  $p$  and  $K$ . They also note that (8) can be interpreted as a mixed effects model with a rank constraint on the covariance matrix. Peng and Paul (2007) propose an alternative geometric based fitting procedure which is somewhat superior to the EM approach.

Figure 2 provides an illustration of this approach on the growth data. The plot gives 80% and 90% confidence intervals for the mean function, and the first two principal components. The confidence intervals were produced using a parametric bootstrap approach (James *et al.*, 2000). Despite the sparsity of the data, the intervals for the mean function are relatively narrow with some widening in the right tail where there are few observations. The confidence intervals for the first principal component are much wider, particularly in the right tail. The large dip in the confidence band in this region occurs because approximately 20% of the bootstrap principal component curves exhibited an inverted U shape. There appear to be two distinctly different possible shapes for this component. Interestingly, given the variability of the first component, the intervals for the second component follow the general shape of the estimated curve quite tightly.

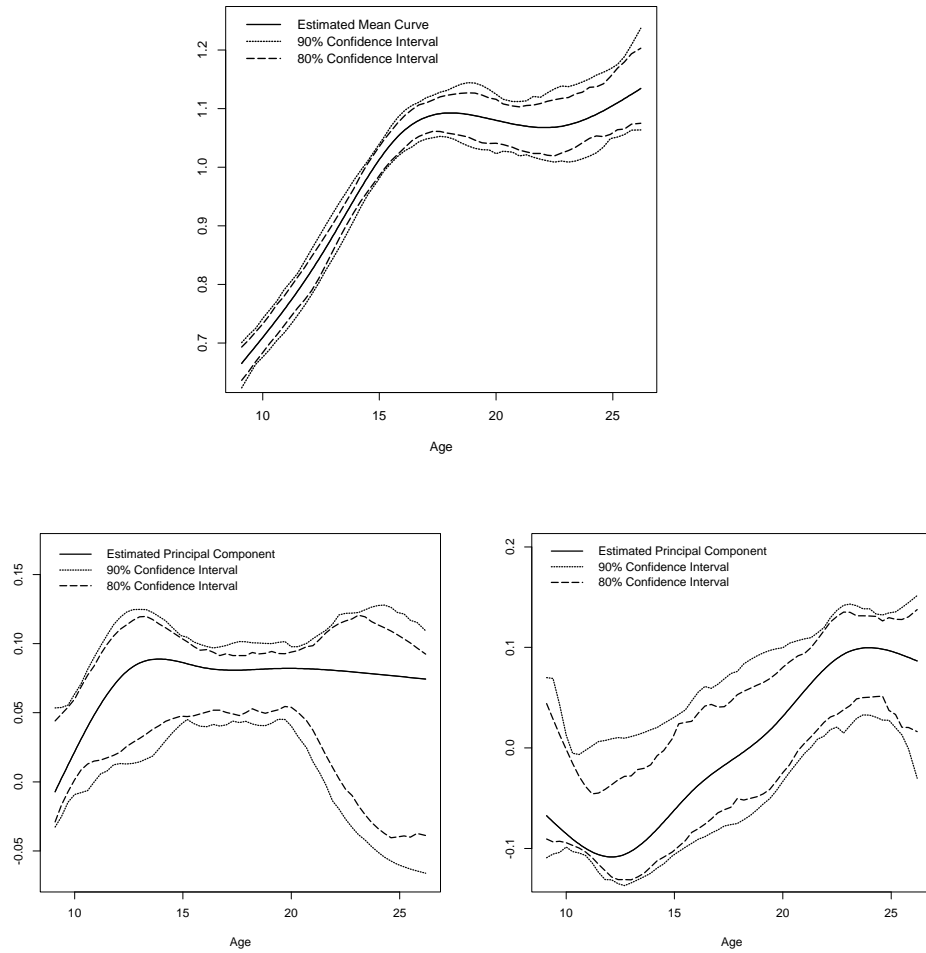


Figure 2: 80% and 90% pointwise confidence intervals for the mean function and both principal components.

## 3.2 Functional Clustering

Here we outline an approach to clustering of sparsely sampled functional data proposed in James and Sugar (2003). In Section 3.2.1 a functional clustering model utilizing the mixed effects framework discussed in Section 2.2 is developed. Section 3.2.2 describes how the model can be used to obtain low-dimensional plots of curve data sets, enabling one to visually assess clustering. Discriminant functions to identify the regions of greatest separation between clusters are developed in Section 3.2.3. Finally, Section 3.2.4 shows an approach for using the clustering model to estimate the entire curve for an individual.

### 3.2.1 A Functional Clustering Model

Let  $g_i(t)$  be the true value for the  $i$ th individual or curve at time  $t$ , and let  $\mathbf{g}_i$ ,  $\mathbf{Y}_i$  and  $\boldsymbol{\epsilon}_i$  be, respectively, the corresponding vectors of true values, observed values and measurement errors at times  $t_{i1}, \dots, t_{in_i}$ . Then  $\mathbf{Y}_i = \mathbf{g}_i + \boldsymbol{\epsilon}_i$ ,  $i = 1, \dots, N$  where  $N$  is the number of individuals. The measurement errors are assumed to have mean zero and to be uncorrelated with each other and  $\mathbf{g}_i$ . It is necessary to impose some structure on the individual curves so, as with the functional PCA approach, one can model  $g_i(t)$  using basis functions. Let  $g_i(t) = \mathbf{s}(t)^T \boldsymbol{\eta}_i$  where  $\mathbf{s}(t)$  is a  $p$ -dimensional spline basis vector and  $\boldsymbol{\eta}_i$  is a vector of spline coefficients. The  $\boldsymbol{\eta}_i$ 's are modeled using a Gaussian distribution,

$$\boldsymbol{\eta}_i = \boldsymbol{\mu}_{\mathbf{z}_i} + \boldsymbol{\gamma}_i, \quad \boldsymbol{\gamma}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma}),$$

where  $\mathbf{z}_i \in \{1, \dots, G\}$  denotes the unknown cluster membership and  $G$  represents the number of clusters. The  $\mathbf{z}_i$ 's can either be modeled as missing data using the ‘‘mixture likelihood’’ approach or as unknown parameters using the ‘‘classification likelihood’’ approach.

There is a further parameterization of the cluster means that proves useful for producing low-dimensional representations of the curves. Note that  $\boldsymbol{\mu}_k$  can be rewritten as

$$\boldsymbol{\mu}_k = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda} \boldsymbol{\alpha}_k, \tag{9}$$

where  $\boldsymbol{\lambda}_0$  and  $\boldsymbol{\alpha}_k$  are respectively  $p$ - and  $h$ -dimensional vectors, and  $\boldsymbol{\Lambda}$  is a  $p \times h$  matrix with  $h \leq \min(p, G - 1)$ . When  $h = G - 1$ , (9) involves no loss of generality while  $h < G - 1$  implies that the means lie in a restricted subspace. With this formulation the functional clustering model (FCM) can

be written as

$$\mathbf{Y}_i = \mathbf{S}_i(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_{\mathbf{z}_i} + \boldsymbol{\gamma}_i) + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, N, \quad (10)$$

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad \boldsymbol{\gamma}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma}),$$

where  $\mathbf{S}_i = (\mathbf{s}(t_{i1}), \dots, \mathbf{s}(t_{in_i}))^T$  is the basis matrix for the  $i$ th curve. Note that  $\boldsymbol{\lambda}_0$ ,  $\boldsymbol{\Lambda}$  and  $\boldsymbol{\alpha}_k$  are confounded if no constraints are imposed. Therefore we require that

$$\sum_k \boldsymbol{\alpha}_k = \mathbf{0} \quad (11)$$

$$\text{and} \quad \boldsymbol{\Lambda}^T \mathbf{S}^T \boldsymbol{\Sigma}^{-1} \mathbf{S} \boldsymbol{\Lambda} = \mathbf{I} \quad (12)$$

where  $\mathbf{S}$  is the basis matrix evaluated over a fine lattice of time points that encompasses the full range of the data and  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I} + \mathbf{S} \boldsymbol{\Gamma} \mathbf{S}^T$ . The restriction in (11) means that  $\mathbf{s}(t)^T \boldsymbol{\lambda}_0$  may be interpreted as the overall mean curve. There are many possible constraints that could be placed on  $\boldsymbol{\Lambda}$ . The reason for the particular form used in (12) will become apparent in Section 3.2.2.

Fitting the functional clustering model involves estimating  $\boldsymbol{\lambda}_0$ ,  $\boldsymbol{\Lambda}$ ,  $\boldsymbol{\alpha}_k$ ,  $\boldsymbol{\Gamma}$  and  $\sigma^2$ . This is achieved by maximizing the likelihood function using an EM algorithm treating the cluster memberships,  $\mathbf{z}_i$ , as missing data. The E-step can be implemented by noting that under (10), conditional on the  $i$ th curve belonging to the  $k$ th cluster,

$$\mathbf{Y}_i \sim N(\mathbf{S}_i(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_k), \boldsymbol{\Sigma}_i) \quad (13)$$

where  $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I} + \mathbf{S}_i \boldsymbol{\Gamma} \mathbf{S}_i^T$ . Depending on whether the classification or mixture likelihood approach is used, curves are first either assigned to a cluster (classification) or assigned a probability of belonging to a cluster (mixture). Then the parameters are estimated given the current assignments and the process is repeated. Further details of the algorithm are provided in James and Sugar (2003).

### 3.2.2 Low-dimensional graphical representations

One of the chief difficulties in high-dimensional clustering is visualization of the data. Plotting functional data is easier because of the continuity of the dimensions. However, it can still be hard to see the clusters since variations in shape and the location of time-points make it difficult to assess the relative distances between curves. These problems are exacerbated when the curves are fragmentary, as in Figure 1a). In this section we illustrate

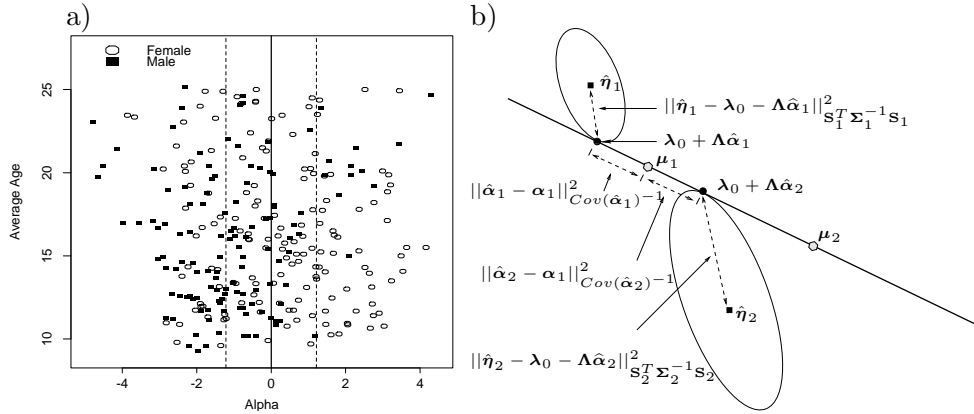


Figure 3: a) A linear discriminant plot for the bone mineral density data. b) An illustration of the decomposition of the distance between two curves and two cluster centers.

a set of graphical tools for use with functional data. The method is based on projecting the curves into a low-dimensional space so that they can be plotted as points, making it much easier to detect the presence of clusters.

Figure 3a) shows the growth data projected onto a one-dimensional space. The horizontal axis represents the projected curve,  $\hat{\alpha}_i$ , while the vertical axis gives the average age of observation for each individual. Points to the left of zero are assigned to cluster 1 and the remainder to cluster 2. Squares represent males and circles females. The dotted lines at  $\alpha_1$  and  $\alpha_2$  correspond to the projected cluster centers. Notice that while there is a significant overlap, most males belong to cluster 1 and most females to cluster 2 even though the model was fit without using gender labels. The plot shows that the clustering separates the genders most strongly for those younger than 16 years. In fact, 74% of such individuals matched the majority gender of their cluster compared with only 57% of those older than 16. This is because girls typically begin their growth spurt before boys.

Figure 3b) illustrates the procedure by which the  $\hat{\alpha}_i$ 's are derived using a two cluster, two curve example. First,  $\mathbf{Y}_i$  is projected onto the  $p$ -dimensional spline basis to get

$$\hat{\eta}_i = (\mathbf{S}_i^T \Sigma_i^{-1} \mathbf{S}_i)^{-1} \mathbf{S}_i^T \Sigma_i^{-1} \mathbf{Y}_i. \quad (14)$$

Second,  $\hat{\eta}_i$  is projected onto the  $h$ -dimensional space spanned by the means,  $\mu_k$ , to get  $\lambda_0 + \Lambda \hat{\alpha}_i$  where

$$\hat{\alpha}_i = (\Lambda^T \mathbf{S}_i^T \Sigma_i^{-1} \mathbf{S}_i \Lambda)^{-1} \Lambda^T \mathbf{S}_i^T \Sigma_i^{-1} \mathbf{S}_i (\hat{\eta}_i - \lambda_0). \quad (15)$$

Thus,  $\hat{\boldsymbol{\alpha}}_i$  is the  $h$ -dimensional projection of  $\mathbf{Y}_i$  onto the mean space after centering. Notice that in this example  $\hat{\boldsymbol{\eta}}_2$  is closest to  $\boldsymbol{\mu}_2$  in Euclidean distance but after projection onto the mean space it is closest to  $\boldsymbol{\mu}_1$  and will be assigned to cluster 1.

James and Sugar (2003) prove that

$$\arg \max_k P(z_{ik} = 1 | \mathbf{Y}_i) = \arg \min_k \left( \|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_k\|_{Cov(\hat{\boldsymbol{\alpha}}_i)^{-1}}^2 - 2 \log \pi_k \right) \quad (16)$$

where

$$Cov(\hat{\boldsymbol{\alpha}}_i) = (\boldsymbol{\Lambda}^T \mathbf{S}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{S}_i \boldsymbol{\Lambda})^{-1} \quad (17)$$

and  $\pi_k$  is the prior probability that an observation belongs to the  $k$ th cluster. From (16) and Bayes rule we note that cluster assignments based on the  $\hat{\boldsymbol{\alpha}}_i$ 's will minimize the expected number of misassignments. Thus no clustering information is lost through the projection of  $\mathbf{Y}_i$  onto the lower dimensional space. We call the  $\hat{\boldsymbol{\alpha}}_i$ 's functional linear discriminants because they are exact analogues of the low-dimensional representations used to visualize data in linear discriminant analysis. In the finite-dimensional setting the linear discriminants all have identity covariance so separation between classes can be assessed visually using the Euclidean distance metric. In the functional clustering setting  $Cov(\hat{\boldsymbol{\alpha}}_i)$  is given by (17). When all curves are measured at the same time points constraint (12) will guarantee  $Cov(\hat{\boldsymbol{\alpha}}_i) = \mathbf{I}$  for all  $i$ , again allowing the Euclidean metric to be used. When curves are measured at different time points it is not possible to impose a constraint that will simultaneously cause  $Cov(\hat{\boldsymbol{\alpha}}_i) = \mathbf{I}$  for all  $i$ . However, when the cluster means lie in a one dimensional subspace ( $h = 1$ ), assuming equal priors, (16) simplifies to

$$\arg \min_k \frac{1}{Var(\hat{\alpha}_i)} (\hat{\alpha}_i - \alpha_k)^2 = \arg \min_k (\hat{\alpha}_i - \alpha_k)^2,$$

which yields the same assignments as if the  $\hat{\alpha}_i$ 's all had the same variance. In this situation it is useful to plot the functional linear discriminants versus their standard deviations to indicate not only to which cluster each point belongs but also the level of accuracy with which it has been observed. Note that for a two cluster model  $h$  must be 1. However, it will often be reasonable to assume the means lie approximately in one dimension even when there are more than two clusters.

Linear discriminant plots have other useful features. Note that the functional linear discriminant for a curve observed over the entire grid of time points used to form  $\mathbf{S}$  will have identity covariance. Thus, the Euclidean

distance between the  $\alpha_k$ 's gives the number of standard deviations separating the cluster means for a fully observed curve. The degree to which the variance for an individual curve is greater than 1 indicates how much discriminatory power has been lost due to taking observations at only a subset of time points. This has implications for experimental design in that it suggests how to achieve minimum variance, and hence optimal cluster separation, with a fixed number of time points. For instance the cluster means in Figure 3a) are 2.4 standard deviations apart, indicating that the groups can be fairly well separated if curves are measured at all time points. The overlap between the two groups is due to the extreme sparsity of sampling, resulting in the  $\hat{\alpha}_i$ 's having standard deviations up to 2.05.

Plots for the membranous nephropathy data, given in Figure 4, provide an example in which the differing covariances of the  $\hat{\alpha}_i$ 's must be taken into account more carefully. Nephrologists' experiences suggest that patients with this disease fall into three groups, either faring well, deteriorating gradually or collapsing quickly. Hence we fit a three cluster model whose mean curves are shown in Figure 4a). With three clusters the means must lie in a plane. Figure 4b) shows a two-dimensional linear discriminant plot with solid circles indicating cluster centers. To circumvent the problem caused by the unequal covariances, we use different symbols for members of different clusters. Note that while most patients fall in the cluster corresponding to their closest cluster in Euclidean distance, there are several that do not. In this example the cluster centers lie essentially on a straight line so it is sufficient to fit a one-dimensional model ( $h = 1$ ). The corresponding plots are shown in Figures 4c) and d). The basic shapes of the mean curves are reassuringly similar to those in 4a), but are physiologically more sensible in the right tail. Figure 4d) plots one dimensional  $\hat{\alpha}_i$ 's versus their standard deviations. We see that the cluster on the right is very tight while the other two are not as well separated. Figures 4e) and f) show respectively the overall mean curve,  $\mathbf{s}(t)^T \boldsymbol{\lambda}_0$  and the function  $\mathbf{s}(t)^T \boldsymbol{\Lambda}$ . The latter, when multiplied by  $\alpha_k$ , gives the distance between  $\mu_k(t)$  and the overall mean curve. From Figure 4e) we see that on average the patients showed a decline in renal function. The primary distinction lies in the speed of the deterioration. For example, the fact that Figure 4f) shows a sharp decline in the first two years indicates that patients in the third cluster, which has a highly positive  $\alpha_3$ , experience a much sharper initial drop than average. In fact all patients in cluster 3 eventually required dialysis.

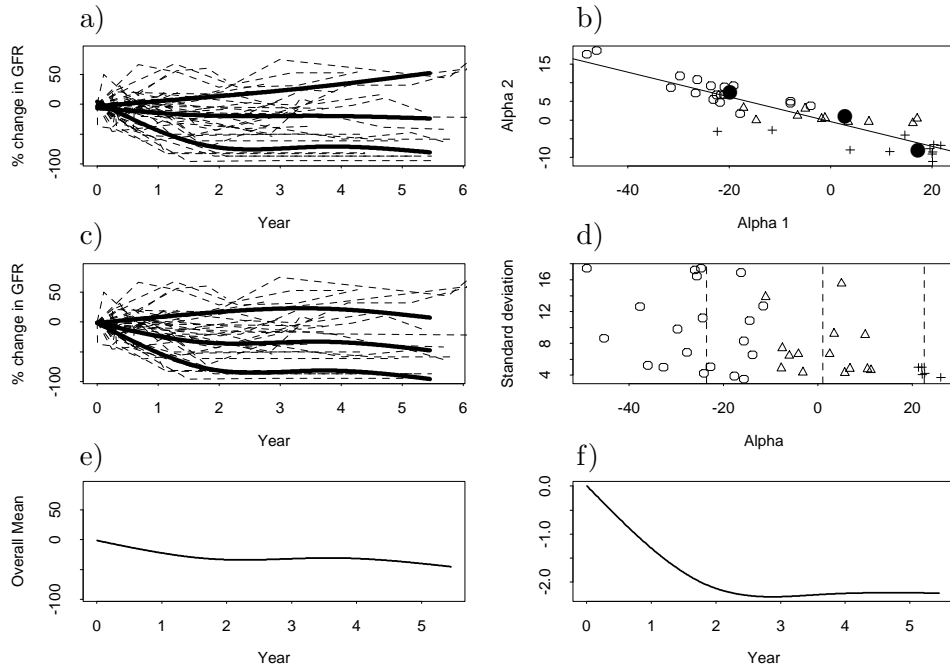


Figure 4: *Assessment plots for the membranous nephropathy data. The cluster mean curves and linear discriminant plots for a fit with  $h = 2$  are shown in a) and b). The equivalent plots for a fit with  $h = 1$  are given in c) and d). Finally, e) shows the overall mean curve and f) the characteristic pattern of deviations about the overall mean.*

### 3.2.3 Discriminant functions

In this section we present a set of curves that identify the dimensions, or equivalently time points, of maximum discrimination between clusters. Intuitively, the dimensions with largest average separation relative to their variability will provide the greatest discrimination. Average separation can be determined by examining  $\mathbf{SA}$  while variability is calculated using the covariance matrix,  $\mathbf{\Sigma} = \mathbf{S}\mathbf{G}\mathbf{S}^T + \sigma^2\mathbf{I}$ . These two quantities can work in opposite directions, making it difficult to identify the regions of greatest discrimination. Consider, for example, Figure 5 which illustrates the covariance and correlation functions for the growth data. From Figure 5a) it is clear that the relationship between a person's bone mineral density before and after puberty is weak but the measurements after puberty are strongly correlated with each other. Figure 5b) has a sharp peak in the early puberty

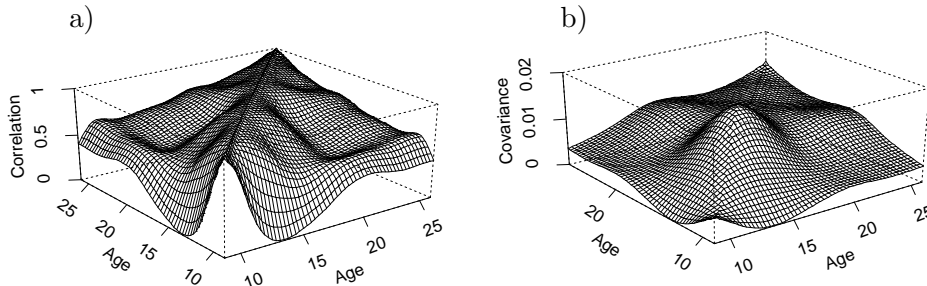


Figure 5: *Estimated a) correlation and b) covariance of  $g_i(t_1)$  with  $g_i(t_2)$ .*

years corresponding to the period of greatest variability. However, this is also the period of greatest distance between the cluster mean curves.

The dimensions of maximum discrimination must also be the ones that are most important in determining cluster assignment. When observations are made at all time points, the spline basis matrix is  $\mathbf{S}$ , and equations (15) and (16) imply that curves should be assigned based solely on the Euclidean distance between  $\hat{\boldsymbol{\alpha}} = \boldsymbol{\Lambda}^T \mathbf{S}^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{S} \boldsymbol{\lambda}_0)$  and the  $\boldsymbol{\alpha}_k$ 's. Thus

$$\boldsymbol{\Lambda}^T \mathbf{S}^T \boldsymbol{\Sigma}^{-1}$$

gives the optimal weights to apply to each dimension for determining cluster membership. Dimensions with low weights contain little information about cluster membership and therefore do little to distinguish among groups, while dimensions with large weights have high discriminatory power. Notice that this set of weights fits with the intuitive notion that dimensions with high discrimination should have large average separation,  $\mathbf{S} \boldsymbol{\Lambda}$ , relative to their variability,  $\boldsymbol{\Sigma}$ . James and Sugar (2003) draw connections between this function and the classical linear discriminant function for a Gaussian mixture.

When the  $\boldsymbol{\alpha}_k$ 's are one dimensional,  $\boldsymbol{\Lambda}^T \mathbf{S}^T \boldsymbol{\Sigma}^{-1}$  is a vector and the weights can be plotted as a single curve, as illustrated by Figure 6 for the growth and nephropathy data sets. For the growth data the highest absolute weights occur in the puberty years, confirming our earlier interpretation from the linear discriminant plot, Figure 3a). For the nephropathy data most of the discrimination between clusters occurs in the early and late

stages of disease. The difference between patients in the later time periods is not surprising. However, the discriminatory power of the early periods is encouraging since one of the primary goals of this study was to predict disease progression based on entry characteristics.

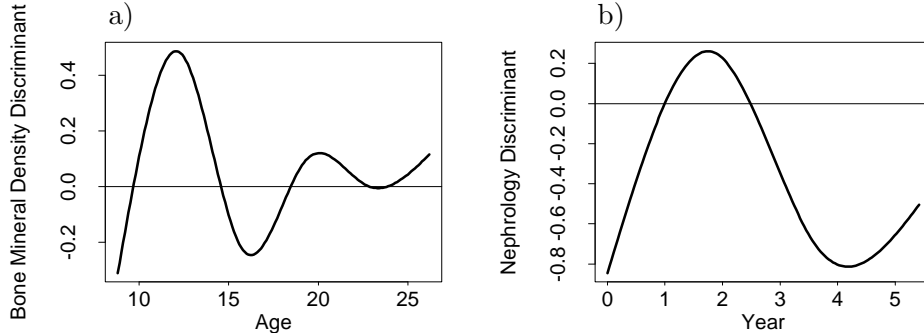


Figure 6: *Discriminant curves for a) the growth data and b) the nephropathy data with  $h = 1$ .*

### 3.2.4 Curve estimation

The model from Section 3.2.1 can also be used to accurately predict unobserved portions of  $g_i(t)$ , the true curve for the  $i$ th individual. When using a basis representation a natural estimate for  $g_i(t)$  is  $\hat{g}_i(t) = \mathbf{s}(t)^T \hat{\boldsymbol{\eta}}_i$ , where  $\hat{\boldsymbol{\eta}}_i$  is a prediction for  $\boldsymbol{\eta}_i$ . James and Sugar (2003) show that the optimal estimate for  $\boldsymbol{\eta}_i$  is  $E(\boldsymbol{\eta}_i | \mathbf{Y}_i)$ . This quantity takes on slightly different values depending on whether the unknown cluster membership,  $\mathbf{z}_i$ , is taken to be a random variable or a parameter to be estimated. However, in the simpler case where  $\mathbf{z}_i$  is modeled as a parameter

$$E(\boldsymbol{\eta}_i | \mathbf{Y}_i) = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda} \boldsymbol{\alpha}_{\mathbf{z}_i} + (\sigma^2 \boldsymbol{\Gamma}^{-1} + \mathbf{S}_i^T \mathbf{S}_i)^{-1} \mathbf{S}_i^T (\mathbf{Y}_i - \mathbf{S}_i (\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda} \boldsymbol{\alpha}_{\mathbf{z}_i})) \quad (18)$$

where  $\mathbf{z}_i = \arg \max_k f(y|z_{ik} = 1)$ . In general, using (18) to form predictions produces significant improvements over the basis approach from Section 2.1 when  $\sigma^2$  is very large, the components of  $\boldsymbol{\Gamma}$  are very small, or  $\mathbf{S}_i^T \mathbf{S}_i$  is close to singular. In fact, when  $\mathbf{S}_i^T \mathbf{S}_i$  is singular the basis approach breaks down completely while the functional clustering method can still produce reliable predictions.

Figure 7 illustrates this approach for two subjects from the growth data. For each plot, the two solid grey lines give the cluster mean curves, the solid black curve fragment gives the observed values for a single individual,

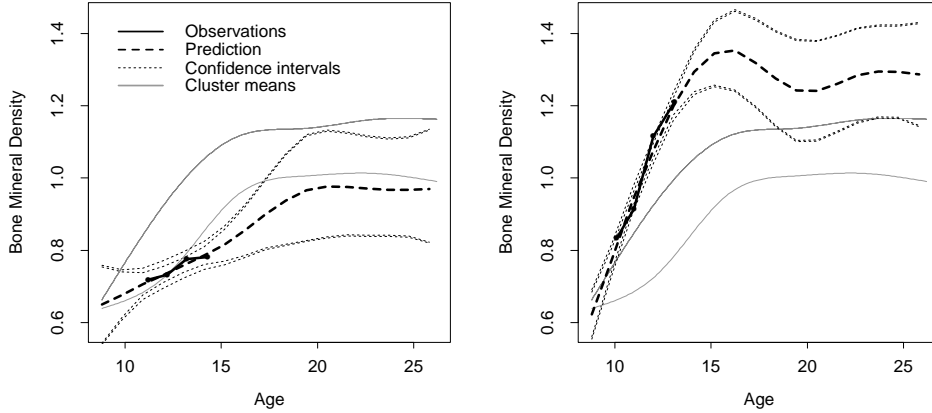


Figure 7: *Curve estimates, confidence intervals, and prediction intervals for two subjects from the growth data.*

and the dashed line gives the corresponding prediction. The dotted lines represent 95% confidence and prediction intervals. See James and Sugar (2003) for details on producing the intervals. Note that the confidence interval provides bounds for the underlying function  $g_i(t)$  while the prediction interval bounds the observed value of  $g_i(t)$ . As usual, the prediction interval is produced by adding  $\sigma^2$  to the variance used in the confidence interval.

### 3.3 Extensions to Functional Classification

The functional clustering model can easily be extended to the functional classification setting where one wishes to classify a curve,  $Y_i(t)$ , into one of  $G$  possible classes. In this setting James and Hastie (2001) use (10) to model the observed curves except that now  $\mathbf{z}_i$  is assumed known for the data used to train the model. This removes one step from the functional clustering EM fitting procedure, namely where  $\mathbf{z}_i$  is estimated. In all other respects the clustering and classification models are fit in an identical fashion.

In addition the same tools that were developed in the clustering setting can also be used for classification problems. For example, the growth data also recorded an individuals ethnicity. We fit the functional classification model to a subset of the data, females of Asian or African American decent, using ethnicity as the class variable. Figure 8a) graphs the corresponding linear discriminants as described in Section 3.2.2. Points to the right of the

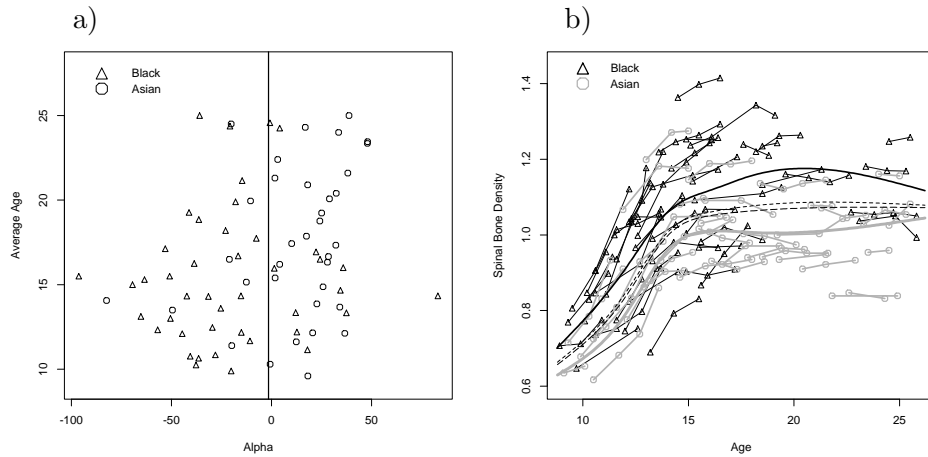


Figure 8: *a) Linear discriminants for African Americans and Asians. The vertical line gives the classification boundary. b) A plot of the raw data for African Americans and Asians. The two solid lines represent the means for African Americans and Asians while the dotted lines are for two other ethnic groups, Hispanics and Whites.*

center vertical line are classified as Asian and those to the left as African American. While there is some overlap between the groups it is clear that most of the observations on the right are Asian (circles) while those on the left tend to be African American (triangles). There is strong evidence of a difference between the two groups. This is further highlighted in Figure 8b) which plots the raw curves for the two ethnic groups along with the estimated mean curves for each group. One can clearly see that African Americans tend to, on average, have higher bone mineral density.

### 3.4 Functional Regression

One of the most useful tools in FDA is that of functional regression. This setting can correspond to either functional predictors or functional responses or both. See Ramsay and Silverman (2002) and Muller and Stadtmuller (2005) for numerous specific applications. One commonly studied problem involves data with functional predictors but scalar responses. Ramsay and Silverman (2005) discuss this scenario and several papers have also been written on the topic, both for continuous and categorical responses, and for linear and non-linear models (Hastie and Mallows, 1993; James and Hastie,

2001; Ferraty and Vieu, 2002; James, 2002; Cardot *et al.*, 2003; Ferraty and Vieu, 2003; Muller and Stadtmuller, 2005; James and Silverman, 2005; Cardot *et al.*, 2007; Crambes *et al.*, 2009). The alternative situation where the response is functional has also been well studied. A sampling of papers examining this situation includes Fahrmeir and Tutz (1994), Liang and Zeger (1986), Faraway (1997), Hoover *et al.* (1998), Wu *et al.* (1998), Fan and Zhang (2000), and Lin and Ying (2001). Chapters 1-6 give good overviews on this topic.

Here we briefly examine an approach by Yao *et al.* (2005b) specifically designed for performing functional regressions where the predictor and response have both been sparsely sampled. Assume we observe a set of predictor curves,  $X_i$ , and a corresponding set of response curves,  $Y_i$ . The predictors are observed at times  $s_{i1}, \dots, s_{il_i}$  and the responses at times  $t_{i1}, \dots, t_{in_i}$ . A standard method to model the relationship between  $X$  and  $Y$  is to use the linear regression model

$$E[Y(t)|X] = \mu_Y(t) + \int \beta(s, t)X^c(s)ds \quad (19)$$

where  $X^c$  is the centered version of  $X$ . In this model,  $\beta(s, t)$  is a two-dimensional coefficient function that must be estimated from the data. When  $X$  and  $Y$  are both densely sampled  $\beta(s, t)$  can be computed relatively simply by using a penalized least squares approach. However, for sparsely observed data the least squares method is not feasible.

Yao *et al.* (2005b) avoid this problem by first assuming that the curves have measurement error using the following model

$$U_{il} = X_i(s_{il}) + \epsilon_{il} = \mu_X(s_{il}) + \sum_{m=1}^{\infty} \zeta_{im}\psi(s_{il}) + \epsilon_{il} \quad (20)$$

$$V_{ij} = Y_i(t_{ij}) + e_{il} = \mu_Y(t_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi(t_{ik}) + e_{ik} \quad (21)$$

Using this model they show that

$$\beta(s, t) = \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \frac{E[\zeta_m \xi_k]}{E[\zeta_m^2]} \psi_m(s) \phi_k(t).$$

Next, scatterplot smoothers are used to obtain smooth estimates of the mean and covariance functions for both the predictors and responses. Then, using a similar approach to that used in Section 2.3, estimates for  $\psi_m(s)$  and

$\phi_k(t)$ , as well as  $\rho_m$ , the eigenvalues for the predictors, can be produced from the covariance functions. A smooth estimate for  $C(s, t) = \text{cov}(X(s), Y(t))$  is also produced using a smoothing procedure. Then  $\rho_m$  can be used to estimate  $E[\zeta_m^2]$  and  $\hat{\sigma}_{km}$  provides an estimate for  $E[\zeta_m \xi_k]$  where

$$\hat{\sigma}_{km} = \int \int \hat{\psi}_m(s) \hat{C}(s, t) \hat{\phi}_k(t) ds dt.$$

The final estimate for  $\beta(s, t)$  is given by

$$\hat{\beta}(s, t) = \sum_{k=1}^K \sum_{m=1}^M \frac{\hat{\sigma}_{km}}{\hat{\rho}_m} \hat{\psi}_m(s) \hat{\phi}_k(t).$$

Once this model has been fit predictions for a new response function,  $Y^*$ , given a partially observed predictor function,  $X^*$ , can be made using

$$Y^*(t) = \hat{\mu}_Y(t) + \sum_{k=1}^K \sum_{m=1}^M \frac{\hat{\sigma}_{km}}{\hat{\rho}_m} \hat{\zeta}_m^* \hat{\phi}_k(t).$$

The only quantity in here that needs to be computed is  $\hat{\zeta}_m^*$ . However,  $\hat{\zeta}_m^*$  can be computed using an analogous approach to the PACE method of Section 2.3. Details are provided in Yao *et al.* (2005b).

## 4 Variable Selection in a Functional Setting

In this section we examine another sense in which sparsity can be important in the regression setting. Here we return to the more traditional FDA paradigm where it is assumed that the functions in question have been observed over a dense grid of time points. In this setting we outline two methods that have been proposed for performing regressions involving a functional predictor,  $X(t)$  and a scalar response,  $Y$ . Both methods use ideas from the high dimensional regression literature, where it is often assumed that the coefficient vector,  $\beta$ , is sparse in the sense of containing few non-zero values.

### 4.1 Selecting Influential Design Points

Ferraty *et al.* (2009) suggest a method, called a “sequential algorithm for selecting design automatically” (SASDA), for taking a function predictor,  $X(t)$ , observed at a fine grid of time points,  $t_1, \dots, t_r$ , and selecting a small

subset of points that are most predictive of the response,  $Y$ . They refer to these points as the “most predictive design points”. The selected design points are then used to produce a local linear regression for predicting the response.

Ferraty *et al.* begin with the general functional nonparametric model

$$Y = g(X) + \epsilon. \quad (22)$$

If one was to interpret this model using a standard nonparametric approach then (22) would become  $Y = g(X(t_1), \dots, X(t_r)) + \epsilon$ . This approach is not practical because  $X(t_1), \dots, X(t_r)$  are both high dimensional and highly correlated. Instead,  $g$  is approximated using a local linear estimator,  $\hat{g}_h$ . Let  $S(\mathbf{t}, h)$  be the cross validated sum of squares error between  $Y_i$  and the prediction,  $\hat{g}_h(X_i)$ , constructed using the time points,  $\mathbf{t}$ . Then an iterative algorithm is proposed for selecting the optimal design points for  $\mathbf{t}$ . First  $\mathbf{t}$  is initialized to the empty set. Then, the single time point  $t_s$  that gives the lowest value for  $S(\mathbf{t}, h)$  is selected. The algorithm continues, at each step selecting the time point resulting in the greatest reduction in  $S(\mathbf{t}, h)$ , conditional on the currently chosen points. This procedure continues until there are no new points that cause a large enough reduction in  $S(\mathbf{t}, h)$ .

Ferraty *et al.* demonstrate SASDA on two real world data sets and show that it can have higher predictive accuracy in comparison to other standard functional linear, or nonlinear, regression methods. Another advantage of SASDA is that it generally selects a small set of time points that jointly contain the relevant information in terms of predicting  $Y$ . This makes the final model much easier to interpret because the relationship between the response and predictor can be summarized in terms of these few time points.

Several extensions of SASDA are also considered. Ferraty *et al.* note that their approach can be computationally expensive. They suggest using a method such as the Lasso (Tibshirani, 1996) to perform an initial dimension reduction and then to implement SASDA on the time points that Lasso selects. This seems to result in a small reduction in prediction accuracy but a significant improvement in computational efficiency. Another extension that is examined involves combining SASDA with other functional regression methods. Specifically, one first fits SASDA and then uses functional linear, or nonlinear, regression methods to predict the leave one out residual between the response and the SASDA prediction. The final predicted response is then the sum of the SASDA fit plus the predicted residual. This combined approach appears to give further improvements in prediction accuracy.

## 4.2 Functional Linear Regression That's Interpretable

James *et al.* (2009) develop a method, which they call “Functional Linear Regression That's Interpretable” (FLiRTI), that produces accurate, but also highly interpretable, estimates for the coefficient function,  $\beta(t)$ . The key to their procedure is to reformulate the problem as a form of variable selection. In particular they divide the time period up into a fine grid of points and then use variable selection methods to determine whether the  $d$ th derivative of  $\beta(t)$  is zero or not at each of the grid points. The implicit assumption is that the  $d$ th derivative will be zero at most grid points i.e. it will be sparse. By choosing appropriate derivatives they can produce a large range of highly interpretable  $\beta(t)$  curves.

James *et al.* start with the standard functional linear regression model

$$Y_i = \beta_0 + \int X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, \dots, n.$$

Then modeling  $\beta(t)$  using a  $p$ -dimensional basis function,  $\beta(t) = \mathbf{B}(t)^T \boldsymbol{\eta}$ , they arrive at

$$Y_i = \beta_0 + \mathbf{X}_i^T \boldsymbol{\eta} + \epsilon_i, \quad (23)$$

where  $\mathbf{X}_i = \int X_i(t)\mathbf{B}(t)dt$ . Estimating  $\boldsymbol{\eta}$  presents difficulties because a high dimensional basis is used so  $p > n$ . Hence FLiRTI models  $\beta(t)$  assuming that one or more of its derivatives are sparse i.e.  $\beta^{(d)}(t) = 0$  over large regions of  $t$  for one or more values of  $d = 0, 1, 2, \dots$ . This approach has the advantage of both constraining  $\boldsymbol{\eta}$  enough to allow one to fit (23) as well as producing a highly interpretable estimate for  $\beta(t)$ . For example, constraining the first derivative produces an estimate for  $\beta(t)$  that is constant over large regions. The fact that the effect of  $X(t)$  on  $Y$  is constant over most time points makes the relationship between predictor and response simple to understand.

Let  $A = [D^d \mathbf{B}(t_1), D^d \mathbf{B}(t_2), \dots, D^d \mathbf{B}(t_p)]^T$  where  $t_1, t_2, \dots, t_p$  represent a grid of  $p$  evenly spaced points and  $D^d$  is the  $d$ th finite difference operator i.e.  $D\mathbf{B}(t_j) = p[\mathbf{B}(t_j) - \mathbf{B}(t_{j-1})]$ ,  $D^2\mathbf{B}(t_j) = p^2[\mathbf{B}(t_j) - 2\mathbf{B}(t_{j-1}) + \mathbf{B}(t_{j-2})]$  etc. Then, if

$$\boldsymbol{\gamma} = A\boldsymbol{\eta}, \quad (24)$$

$\gamma_j$  provides an approximation to  $\beta^{(d)}(t_j)$  and hence, enforcing sparsity in  $\boldsymbol{\gamma}$  constrains  $\beta^{(d)}(t_j)$  to be zero at most time points. For example, one may believe that  $\beta^{(2)}(t) = 0$  over many regions of  $t$ , i.e.  $\beta(t)$  is exactly linear over large regions of  $t$ . In this situation we would let

$$A = [D^2\mathbf{B}(t_1), D^2\mathbf{B}(t_2), \dots, D^2\mathbf{B}(t_p)]^T \quad (25)$$

which implies  $\gamma_j = p^2[\mathbf{B}(t_j)^T \boldsymbol{\eta} - 2\mathbf{B}(t_{j-1})^T \boldsymbol{\eta} + \mathbf{B}(t_{j-2})^T \boldsymbol{\eta}]$ . Hence, provided  $p$  is large, so  $t$  is sampled on a fine grid,  $\gamma_j \approx \beta^{(2)}(t_j)$ . In this case enforcing sparsity in the  $\gamma_j$ 's will produce an estimate for  $\beta(t)$  that is linear except at the time points corresponding to non-zero values of  $\gamma_j$ .

If  $A$  is constructed using a single derivative, as in (25), then one can always choose a grid of  $p$  different time points,  $t_1, t_2, \dots, t_p$  such that  $A$  is a square  $p$  by  $p$  invertible matrix. In this case  $\boldsymbol{\eta} = A^{-1}\boldsymbol{\gamma}$  so (23) and (24) can be combined to produce the FLiRTI model

$$\mathbf{Y} = V\boldsymbol{\gamma} + \epsilon$$

where  $V = [\mathbf{1}|XA^{-1}]$ ,  $\mathbf{1}$  is a vector of ones and  $\beta_0$  has been incorporated into  $\boldsymbol{\gamma}$ . James *et al.* then use high dimensional regression methods such as the Lasso (Tibshirani, 1996) and the Dantzig Selector (Candes and Tao, 2007) to estimate  $\boldsymbol{\gamma}$ . A final estimate for  $\beta(t)$  is produced using  $\hat{\beta}(t) = \mathbf{B}(t)^T \hat{\boldsymbol{\eta}} = \mathbf{B}(t)^T A^{-1} \hat{\boldsymbol{\gamma}}$ . James *et al.* show that FLiRTI performs well on simulated and real world data sets. In addition, they present non-asymptotic theoretical bounds on the estimation error.

## References

- BACHRACH, L. K., HASTIE, T. J., WANG, M. C., NARASIMHAN, B., AND MARCUS, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth; a longitudinal study. *Journal of Clinical Endocrinology & Metabolism*, **84**, 4702–4712.
- BANFIELD, J. D. AND RAFTERY, A. E. (1993). Model-based Gaussian and non-gaussian clustering. *Biometrics*, **49**, 803–821.
- BRUMBACK, B. AND RICE, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961–976.
- CANDES, E. AND TAO, T. (2007) The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *Annals of Statistics*, **35**, 2313–2351.
- CARDOT, H., FERRATY, F., AND SARDA, P. (2003) Spline estimators for the functional linear model. *Statist. Sinica*, **13**, 571–591.
- CARDOT, H., MAS, A., AND SARDA, P. (2007) CLT in functional linear regression models. *Probab. Theory Related Fields*, **138**, 325–561.
- CRAMBES, C., KNEIP, A., AND SARDA, P. (2009) Smoothing splines estimators for functional linear regression. *Ann. Statist.*, **37**, 35–72.
- DIPILLO, P. J. (1976). The application of bias to discriminant analysis. *Communications in Statistics, Part A - Theory and Methods*, **A5**, 843–854.

- FAHRMEIR, L. AND TUTZ, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer-Verlag.
- FAN, J. AND ZHANG, J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B*, **62**, 303–322.
- FARAWAY, J. (1997). Regression analysis for a functional response. *Technometrics*, **39**, 254–261.
- FERRATY, F., HALL, P., AND VIEU, P. (2009). Most-predictive design points for functional data predictors. *Biometrika*, (to appear).
- FERRATY, F. AND VIEU, P. (2002). The functional nonparametric model and applications to spectrometric data. *Computational Statistics*, **17**, 545–564.
- FERRATY, F. AND VIEU, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis*, **44** 161–173.
- FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165–175.
- GREEN, P. J. AND SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models : A roughness penalty approach*, Chapman and Hall: London.
- HASTIE, T. AND MALLOW, C. (1993). Comment on “a statistical view of some chemometrics regression tools”. *Technometrics*, **35**, 140–143.
- HASTIE, T. J., BUJA, A., AND TIBSHIRANI, R. J. (1995). Penalized discriminant analysis. *Annals of Statistics*, **23**, 73–102.
- HENDERSON, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics*, **21**, 309–310.
- HOOVER, D. R., RICE, J. A., WU, C. O., AND YANG, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.
- JAMES, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, **64**, 411–432.
- JAMES, G. M. AND HASTIE, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B*, **63**, 533–550.
- JAMES, G. M., HASTIE, T. J., AND SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.
- JAMES, G. M. AND SILVERMAN, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association*, **100**, 565–576.
- JAMES, G. M. AND SUGAR, C. A. (2003). Clustering for sparsely sampled

- functional data. *Journal of the American Statistical Association*, **98**, 397–408.
- JAMES, G. M., WANG, J., AND ZHU, J. (2009). Functional Linear Regression That’s Interpretable. *Annals of Statistics*, **37**, 2083–2108.
- LAIRD, N. M. AND WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
- LIANG, K. Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
- LIN, D. Y. AND YING, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, **96**, 103–113.
- MULLER, H. G. AND STADTMULLER, U. (2005). Generalized functional linear models. *Annals of Statistics*, **33**, 774–805.
- PENG, J. AND PAUL, D. (2007). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *University of California-Davis, Technical Report*.
- RAMSAY, J. O. AND SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*. Springer.
- RAMSAY, J. O. AND SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer, 2nd edn.
- RICE, J. A. AND WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.
- SHI, M., WEISS, R., AND TAYLOR, J. (1996). An analysis of pediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics*, **45**, 151–164.
- SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *R. Statist. Soc. B*, **47**, 1–52.
- STANISWALIS, J. G. AND LEE, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, **93**, 1403–1418.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Roy. Statist. Soc. Ser. B*, **58**, 267–288.
- WU, C. O., CHIANG, C. T., AND HOOVER, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, **93**, 1388–1402.
- YAO, F., MULLER, H., AND WANG, J. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.

YAO, F., MULLER, H., AND WANG, J. (2005b). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, **33**, 2873–2903.