Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves

John A. Rice
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720

Colin O. Wu
Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, MD 21218

June 17, 1998

Abstract

We propose a method of analyzing collections of related curves in which the individual curves are modeled as spline functions with random coefficients. The method is applicable when the individual curves are sampled at variable and irregularly spaced points. This produces a low rank, low frequency approximation to the covariance structure, which can be estimated naturally by the EM algorithm. Smooth curves for individual trajectories are constructed as BLUP estimates, combining data from that individual and the entire collection. This framework leads naturally to methods for examining the effects of covariates on the shapes of the curves. We use model selection techniques—AIC, BIC, and cross-validation—to select the number of breakpoints for the spline approximation. We believe that the methodology we propose provides a simple, flexible, and computationally efficient means of functional data analysis. We illustrate it with two sets of data.

1 Introduction

In recent years there has been an increasing interest in non-parametric analysis of data that is in the form of noisy sampled points from collections of curves. Methodology focusing on the curves themselves as the objects of interest has come to be known as "functional data analysis" (Ramsay and Silverman, 1997). Noteworthy early work in functional data analysis includes non-parametric analysis of growth curves (Gasser, Müller, Köhler, Molinari and Prader, 1984). The aims of the analysis include smoothing individual curves, estimating functionals of individual curves, examining covariate effects on the shapes of the curves, and the decomposition of each curve into a mean function and a few major modes of variability. In this paper we present a methodology for analysis of collections of curves that may be unequally and sparsely sampled.

We extend current methodology for analyzing repeated measures data via linear mixed effects models to a nonparametric setting. The formulation of Laird and Ware (1982) was a key development; general expositions are provided in Diggle, Liang and Zeger (1994), Jones (1993) and Vonesh and Chincilli (1977). A typical such mixed effects analysis represents each subject's repeated measures as the sum of a population mean function depending on time and other covariates, a low degree polynomial with random coefficients, and white measurement noise. The white noise model is sometimes broadened to include a stationary continuous time process, such as the Ornstein-Uhlenbeck process, to account for autocorrelation;

the process is usually chosen rather arbitrarily for convenience. The polynomial random effects and the white noise or stationary process thus determine the covariance structure. The EM algorithm is typically used for estimation in these mixed effects models (Laird and Ware, 1982).

In this paper we model the individual curves as spline functions with random coefficients, consequently approximating the covariance function as a tensor product of splines. This produces a low rank, low frequency approximation to the covariance structure like that accomplished via eigenfunction decomposition (Rice and Silverman, 1991), without requiring the data to be regularly spaced and without an artificial imposition of stationarity. The model also includes white measurement noise. Estimation of the covariance structure is accomplished naturally through the EM algorithm, and the resulting covariance kernel can be decomposed into eigenfunctions. Smooth curves for individual trajectories are constructed as BLUP estimates (Robinson, 1991), combining data from that individual and the entire collection. Within this framework it is simple to examine covariate effects. We use model selection techniques—AIC, BIC, and cross-validation—to select the number of knots of the splines. We believe that the methodology we propose provides a simple, flexible, and computationally efficient means of functional data analysis.

The method is related to that of Brumback and Rice (1997), who also use splines, but assume a particular form of the covariance kernel arising from smoothing splines. As in Stone, Hansen, Kooperberg and Truong (1997) we have found it convenient to use splines with a small number of knots, but other bases could be used as well. In Besse, Cardot and Ferraty (1997), splines with a small number of knots were used to smooth individual curves in a quite different way. In particular, we do not fit each curve separately—indeed the data from an individual subject may be too sparse to support such a fit.

The remainder of the paper is organized as follows: in section 2 we present the general methodology. Section 3 contains applications to the gait data analyzed in Rice and Silverman (1991) and an analysis of a sequence of CD4 counts similar to that presented in Zeger and Diggle (1994). Section 4 contains some closing comments.

2 Methodology

Let there be m subjects, n_i observations at times $0 \leq t_{ij} \leq T$ on the i-th subject, and $n = \sum_{i=1}^{m} n_i$ observations in all. Let $Y_{ij} = Y_i(t_{ij})$ be the outcome measured on the i-th subject at time t_{ij}. To keep things simple initially, suppose that there are no covariates other than time. The time series of measurements of an individual subject is represented as the sum of a population mean function, a random function, and white noise. We model the mean function and the random function with splines. The mean function is

$$E(Y_i(t)) = \mu(t) = \sum_{k=1}^{p} \beta_k \bar{B}_k(t), \tag{1}$$

where $\{\bar{B}_k()\}$ is a basis for spline functions on $[0, T]$ with a fixed knot sequence (in our computations we use the B-spline basis and equally spaced knots). The random effect curve for the i-th subject is similarly modeled as the spline function $\sum_{k=1}^{q} \gamma_{ik} B_k(t)$. Here $\{B_k()\}$ is a basis for a possibly different space of spline functions on $[0, T]$ and the γ_{ik} are random coefficients with mean 0 and covariance matrix Γ. Finally, incoporating uncorrelated measurement errors ϵ_{ij} with mean zero and constant variance σ^2, our model is

$$Y_{ij} = \sum_{k=1}^{p} \beta_k \bar{B}_k(t_{ij}) + \sum_{k=1}^{q} \gamma_{ik} B_k(t_{ij}) + \epsilon_{ij} \tag{2}$$

The approximate covariance kernel for a random curve $Y(t)$ is thus modeled nonparametrically as

$$Cov(Y(s), Y(t)) = \sum_{k=1}^{q} \sum_{l=1}^{q} \Gamma_{kl} B_k(s) B_l(t) + \sigma^2 \delta(s - t), \tag{3}$$

where $\delta()$ is the Dirac delta function. Viewing this as an approximation, low frequency components of the covariance kernel are captured in the first term and the remainder is approximated by the second term.

Now (2) is a classical linear mixed effects model and the vector of observations on the i-th subject can be expressed as

$$Y_i = X_i \beta + Z_i \gamma_i + \epsilon_i. \tag{4}$$

The covariance matrix of Y_i is $V_i = Z_i \Gamma Z_i^T + \sigma^2 I$. We can thus use the methodology that has been developed for mixed effect models in this nonparametric context. Estimation of the parameters β, σ^2, and the covariance matrix Γ is accomplished by the EM algorithm (Laird and Ware, 1982). The BLUP estimate (Robinson, 1991) of the spline coefficients of the random effect for subject i is

$$\hat{\gamma}_i = \hat{\Gamma} Z_i^T (Z_i \hat{\Gamma} Z_i^T + \hat{\sigma}^2 I)^{-1} (Y_i - X_i \hat{\beta}). \tag{5}$$

The corresponding estimate of an individual trajectory is then the smooth curve

$$\hat{Y}_i(t) = \sum_{k=1}^{p} \hat{\beta}_k \bar{B}_k(t) + \sum_{k=1}^{q} \hat{\gamma}_{ik} B_k(t). \tag{6}$$

This estimate combines information from the entire sample and from the individual subject in that it uses the population covariance structure to estimate the spline coefficients and shrinks the curve toward the population mean. We note that this estimate is well defined even when the observations on a particular subject are too sparse to support an ordinary least squares fit.

There is not a simple analytic connection between the covariance matrix Γ and the eigenstructure of the covariance kernel (3). However the first term of (3) can be evaluated on a fine grid using the estimate $\hat{\Gamma}$, and the eigenvectors of the resulting matrix can be evaluated numerically. The projection of Y_i on a particular eigenfunction can be determined by evaluating (6) on the same grid and then forming the inner product with the corresponding eigenvector. It can be useful to plot these scores against each other or against covariates.

An alternative form of eigenanalysis is based on the eigenvectors of Γ. The trajectory (6) is then decomposed by expressing $\hat{\gamma}_i$ as a linear combination of those eigenvectors.

For practical application, the number and locations of the knots for the splines corresponding to the mean function and the random effects have to be specified. We have not found this difficult to establish in a seemingly satisfactory way, since in the examples we have examined so far the results are rather insensitive to the specification; a relatively small number of equispaced knots has generally been sufficient. For objective guidance we have resorted to model selection criteria. Specifically, we have cross validated the Gaussian log likelihood, which is the sum of the contributions from the individual curves:

$$\ell_i = -\frac{n_i}{2} \log(2\pi) - \frac{n_i}{2} \log\det V_i - \frac{1}{2}(Y_i - \mu_i)^T V_i^{-1} (Y_i - \mu_i) \tag{7}$$

Here, $\mu_i = (\mu(t_{i1}), \ldots, \mu(t_{in_i}))^T$. The cross-validated log likelihood finds a balance between the last term above, which decreases with complexity of the covariance function, and the second term, which increases. In the examples which follow, we also use AIC and BIC which give results qualitatively comparable to those obtained by cross validation, and are faster to compute.

We now discuss the incorporation of covariates. First, fixed effects can be included by adding columns to the design matrices X_i in the usual way. More interestingly, our framework provides for the examination of time-varying effects of time-independent random covariates in a natural way. Denoting such a covariate by U, from (2)

$$E(Y_i(t)|U = u) = \mu(t) + \sum_{k=1}^{q} E(\gamma_{ik}|U = u)B_k(t) + E(\epsilon_i(t)|U = u). \tag{8}$$

For a linear model,

$$E(\gamma|U = u) = \Sigma_{\gamma U}\Sigma_{UU}^{-1}(u - E(U)), \tag{9}$$
$$E(\epsilon|U = u) = \Sigma_{\epsilon U}\Sigma_{UU}^{-1}(u - E(U)). \tag{10}$$

The covariance matrices can be estimated from the data. For example, the natural estimate of $\Sigma_{\gamma U}$ is

$$S_{\gamma U} = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i(u_i - \bar{u})^T \tag{11}$$

where $\hat{\gamma}_i$ is given by (5). In this paper we only consider such linear estimates for $E(\gamma_{ik}|U = u)$, but the possibility of using more sophisticated methods for predicting the random spline coefficients by covariates is evident. In the case that the covariate is categorical, this analysis simply amounts to averaging the spline coefficients and the "errors," ϵ , at each level of the covariate.

3 Examples

In this section we treat two examples. The first is that of Rice and Silverman (1991) on human gait and the second is of time histories of CD4 counts.

3.1 Human Gait

These data were collected by the Motion Analysis Laboratory of Children's Hospital, San Diego and were used to illustrate eigenfunction analysis in Rice and Silverman (1991). Full details are given in Olshen, Biden, Wyatt and Sutherland (1989). Here we consider curves formed by the angles of the hip over gait cycles of 39 children. Time is measured as a fraction of each individual's gait cycle beginning and ending at the point at which the heel strikes the ground, and the numbers of observations ranged from 16 to 22 per cycle. In Rice and Silverman (1991), the data were interpolated to give 20 equispaced points per cycle, a procedure which seemed reasonable, especially in light of the large signal to noise

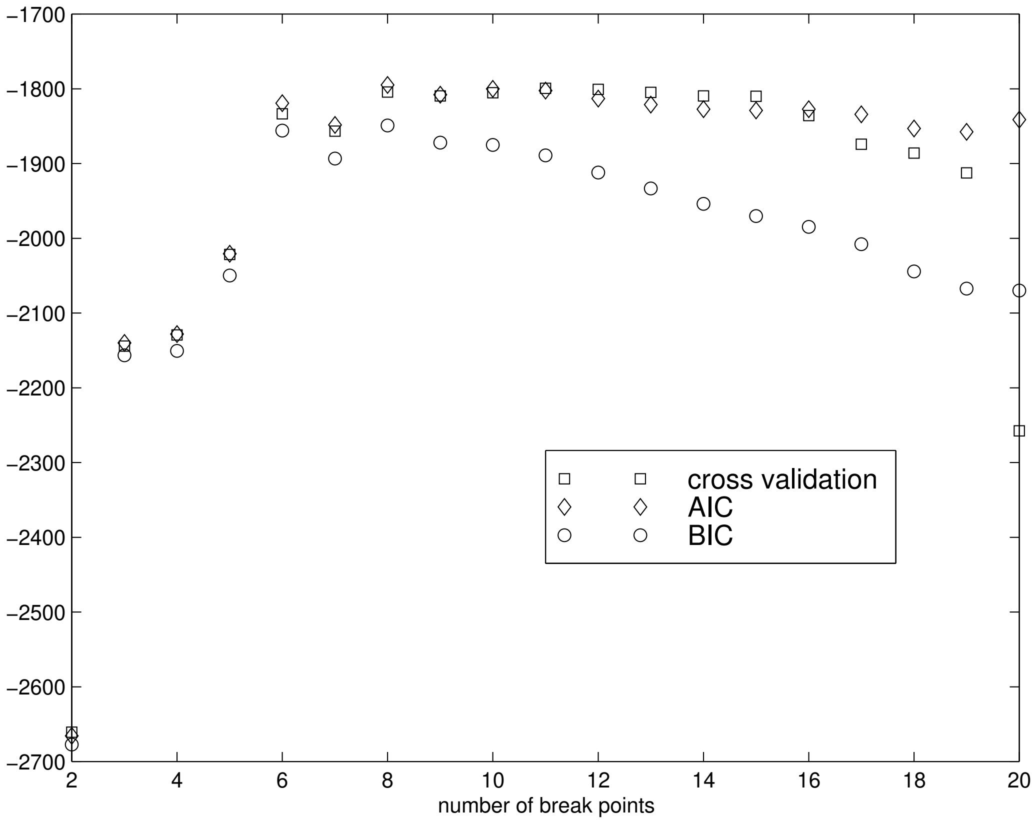

Figure 1: Model selection criteria for the number of breakpoints for the hip angle curves.

ratio. Here we do not interpolate, but use the methods outlined in the previous section.

Figure 1 shows the cross validated log likelihood, the AIC criterion, and the BIC criterion of cubic splines for the mean function and random effects. For simplicity we restricted them to have the same number of equally spaced breakpoints; we count breakpoints to include those at 0 and 1, so a pure cubic has two breakpoints, for example. The cross validation function and the AIC criterion are rather flat in the region from six to about fifteen breakpoints, whereas the BIC criterion drops more rapidly after reaching a maximum at eight breakpoints. The high signal to noise ratio supports a fairly high dimensional approximation (we also added noise to the data and indeed the criteria then peaked at lower dimensions).

The covariance function estimated from ten breakpoints is shown in Figure 2, which shows high variability during the early and late cycle and strong correlation between angles at these times. The estimated mean and the first two eigenfunctions of the covariance function displayed in Figure 2 are shown in Figure 3 for three, ten and twenty breakpoints. Under and over-smoothing is visually apparent in the extreme cases, but qualitatively the results are strikingly similar. The results for ten breakpoints is very similar to those obtained in Rice and Silverman (1991) after interpolation. The first eigenfunction represents deviations from the mean curve which are of the same sign throughout the cycle, with increased amplitude at the beginning and end of the cycle. The second eigenfunction portrays a shift in amplitude which is not constant throughout the cycle. Here the major effect of over-fitting is roughness in the estimated eigenfunctions and in the corresponding covariance function near the boundaries. Other than this roughness, the first eigenfunction (the smoothest) is estimated fairly consistently by the three smoothings.

Choices of ten or twenty breakpoints give very similar estimates of the mean curve whereas there is apparently noticeable systematic distortion when three are used.

Finally, Figure 4 shows the BLUP smoothing (6) of a single curve for three, ten, and twenty breakpoints. The most striking feature of this figure is the jagged curve near the endpoints produced by overfitting with 20 breakpoints. This is due to the contribution from the eigenfunctions as shown in Figure 3 and discussed above. The choice of three breakpoints oversmooths the data.

3.2 CD4 Counts

As a second example we consider sequences of CD4 counts from 463 homosexual men from the Multi-Center AIDS Cohort Study who seroconverted between 1984 and 1993; see Kaslow, Ostrow, Detels, Phair, Polk and Rinaldo (1987) for details of the study design and methods. Since HIV destroys CD4 cells, their counts are a standard method of measuring disease progression. In contrast to the previous example, individual curves are sparsely and irregularly sampled. The number of observations per subject ranged from 1 to 16 over follow up periods ranging up to 94 months after becoming HIV positive. In contrast to the gait data, the observations are noisy and sparse and, other than a decreasing trend, dynamical features are not visually apparent. Covariates include age at the time of seroconversion and smoking status (recorded as ever smoked or never smoked during the study, and hence time-independent). Similar data were analyzed by Zeger and Diggle (1994), using a semi-parametric model, and by Fan and Zhang (1997) to illustrate

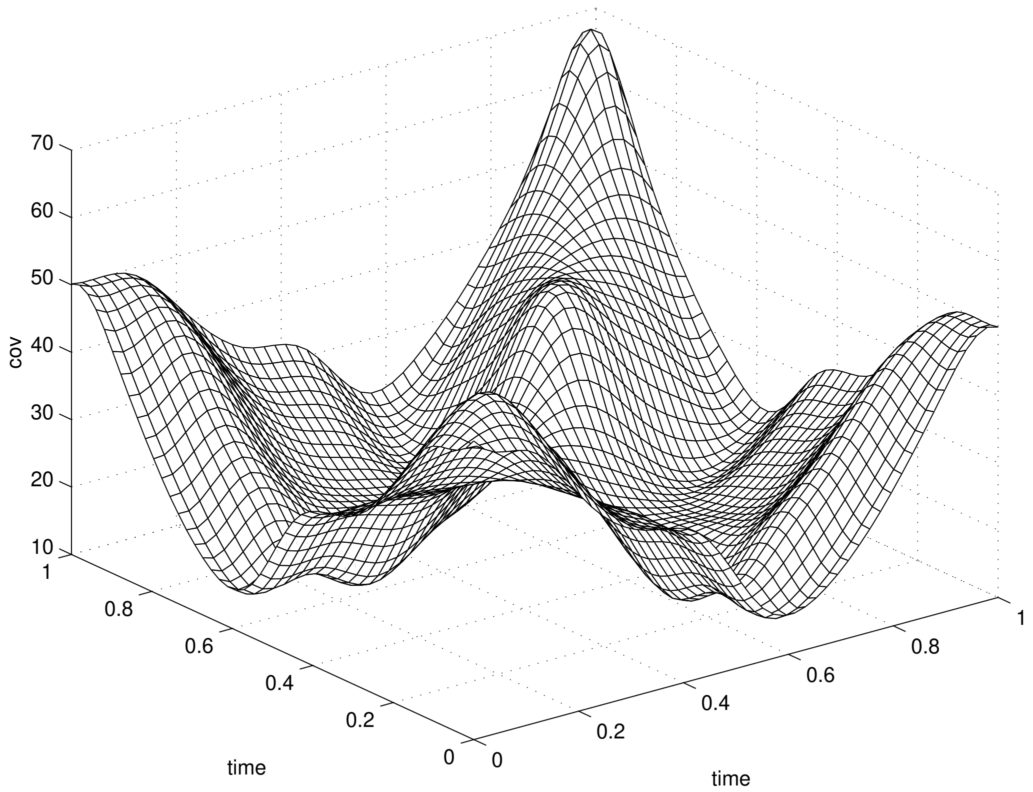

Figure 2: Covariance function for a hip cycle estimated with ten breakpoints.

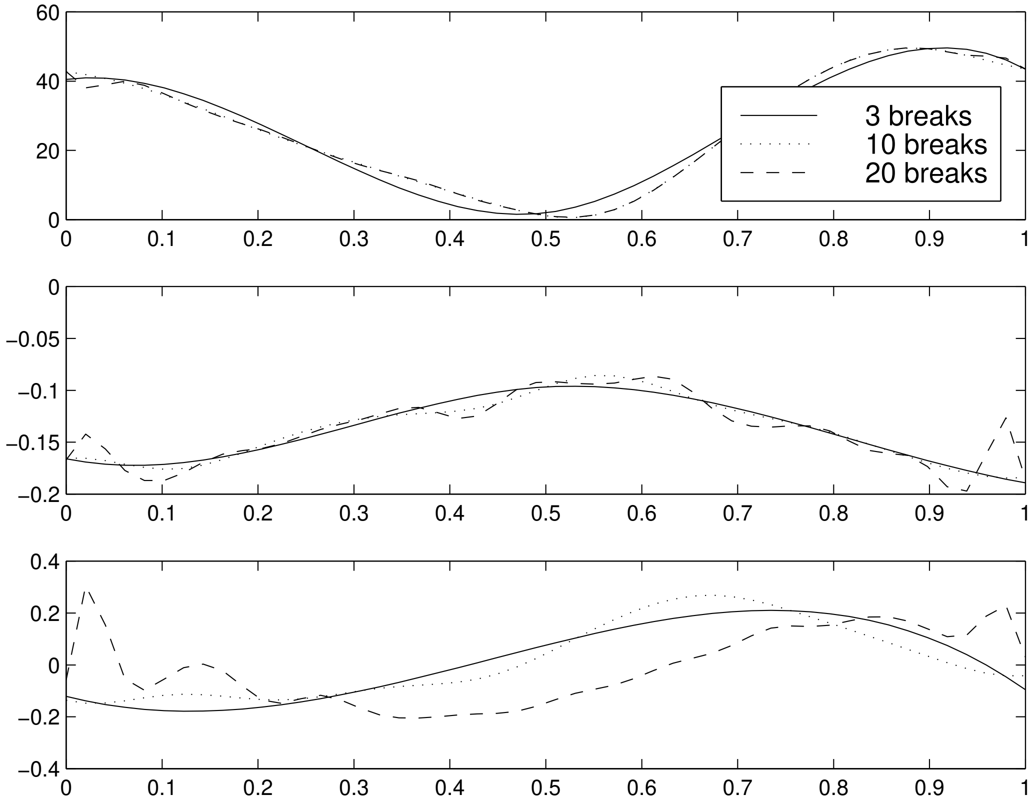

Figure 3: Estimates of the mean function (top panel) and the first two eigenfunctions of a hip cycle by splines with three, ten, and twenty breakpoints.

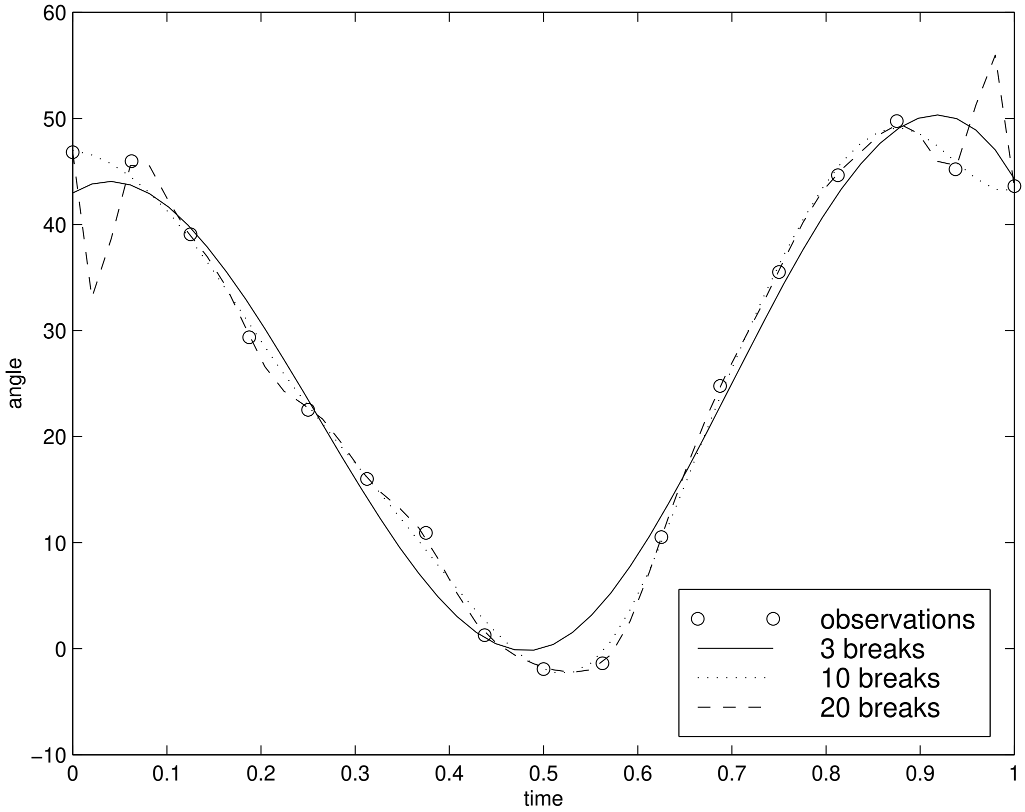

Figure 4: Three smoothings of the data from a single individual.

the use of a particular functional linear model.

Based on AIC and BIC scores, we used four equally spaced breakpoints for cubic spline functions for the mean and random effects. The mean function is shown in Figure 5 along with the individual trajectories. The covariance function is shown in Figure 6; it is clearly non-stationary with high variability at early and late times. Variability at early times may be in part due to lack of precision in identifying the actual date of seroconversion.

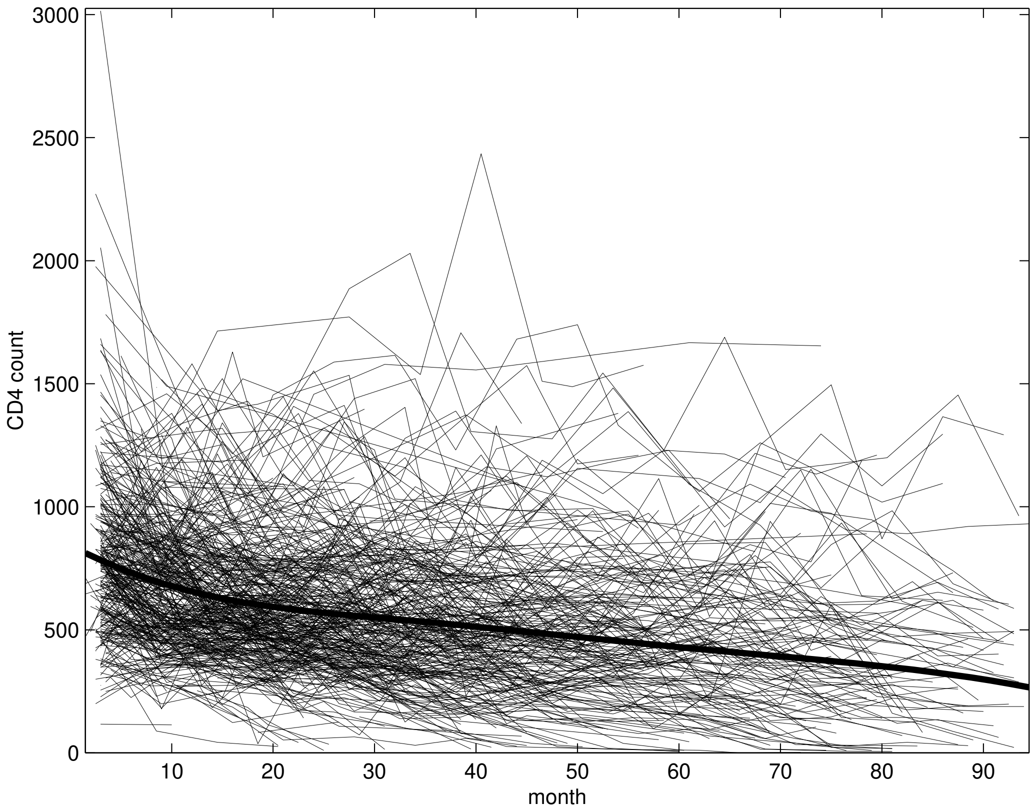

Figure 5: Estimated mean function and individual trajectories of 463 sequences of CD4 counts

It can be useful to single out unusual individual cases, but this can be difficult when it is not *a priori* clear what is meant by "unusual" and when direct visual examination of the data is complicated by irregular sampling, substantial noise, and a large number of curves. Eigenfunction analysis of the covariance kernel (3) provides one possible approach, but here we consider another, based on the

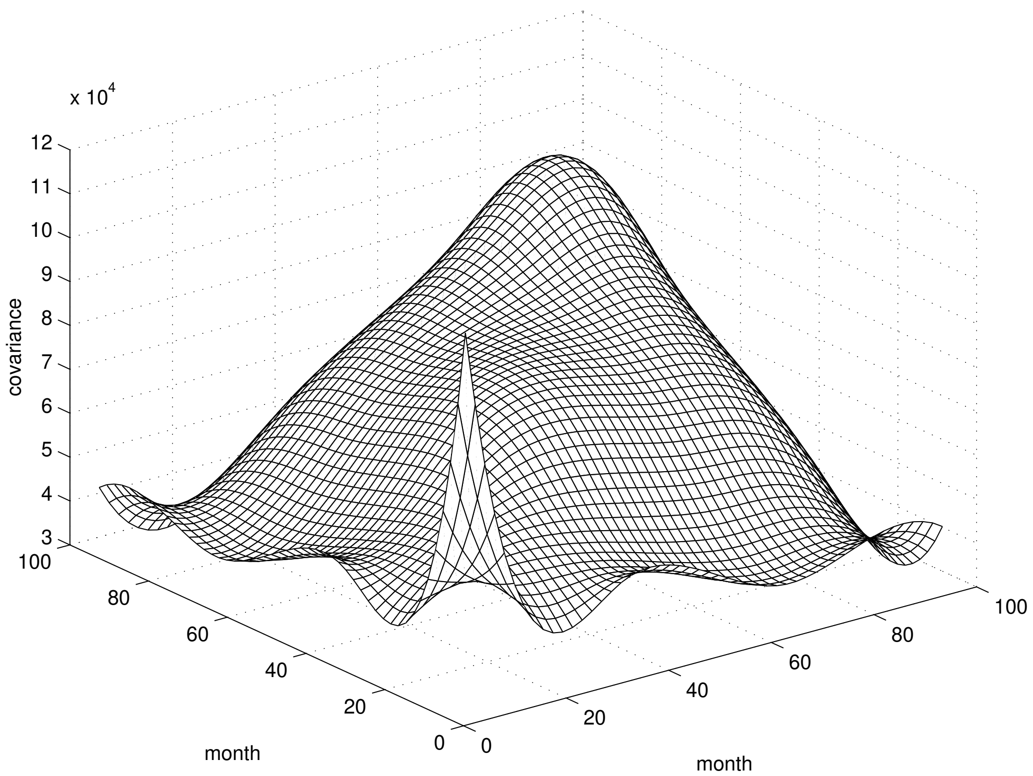

Figure 6: Estimated covariance function of CD4 counts

BLUP estimates of the random spline coefficients, $\hat{\gamma}_i$, (5) and their estimated covariance matrix $\hat{\Gamma}$. We first consider choosing individual trajectories with large Mahalanobis distances, $\hat{\gamma}^T\hat{\Gamma}^{-1}\hat{\gamma}$. Figure 7 shows the measurements of two such subjects along with their BLUP smooths. These subjects are unusual because their CD4 counts are high overall (compare to Figure 5).

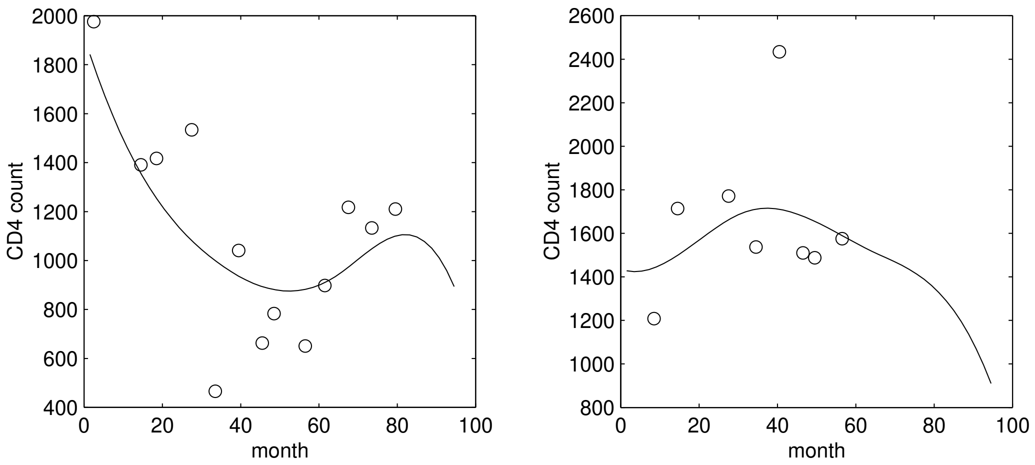

Figure 7: CD4 counts for two subjects with large Mahalanobis distances and the corresponding BLUP smooths

We next consider the eigenfunction decomposition of $\hat{\Gamma}$, shown in Figure 8. Since the covariance matrix is that of spline coefficients with respect to a B-spline basis, the eigenvectors are difficult to interpret directly, so we show the splines corresponding to them. That is, if ξ is an eigenvector, we display the "eigenspline" $\sum \xi_k B_k(t)$. (Note that this eigendecomposition is an alternative to that shown for the gait data above, where we found the eigenfunctions of the covariance function (3).) The first eigenvector corresponds to an overall shift of CD4 level, the second to a trend which is especially steep in the early months, the third to an initial increase followed by a downward trend, and the fourth by a reversal of trend at about month 40. Figure 9 is a scatter plot of the scores of the cases on the second and third eigenvector, the inner products of the BLUP estimates of the coefficients with the respective eigenvectors. Figure 10 shows unusual cases as defined by extreme scores. The first is unusual in having a high initial value followed by a very rapid decrease. Unlike most of the cases, the second peaks at about 30 months and then decreases to a low level.

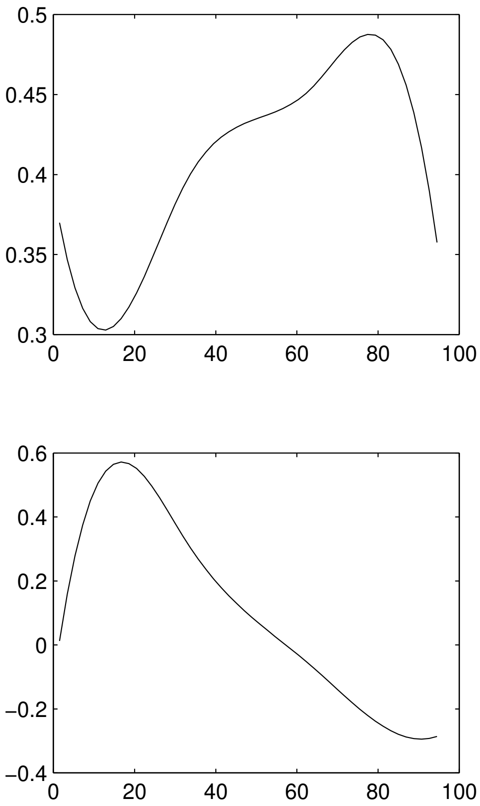

Figure 8: Eigensplines of CD4 count curves, ordered left to right and across rows.

We examined the effects of age and smoking status. We standardized subjects' ages, which ranged from 18 to 64. Figure 11 shows the covariate "effect" curve (8) resulting from modeling the dependence on age linearly as in (9) and (10). (The assumption of linearity was informally checked by plotting age versus BLUP estimates of spline coefficients.) Also shown are error bars found by the bootstrap as in Hoover, Rice, Wu and Yang (1998) (subjects were sampled with replacement 100 times; the error bars are the pointwise standard deviations of the

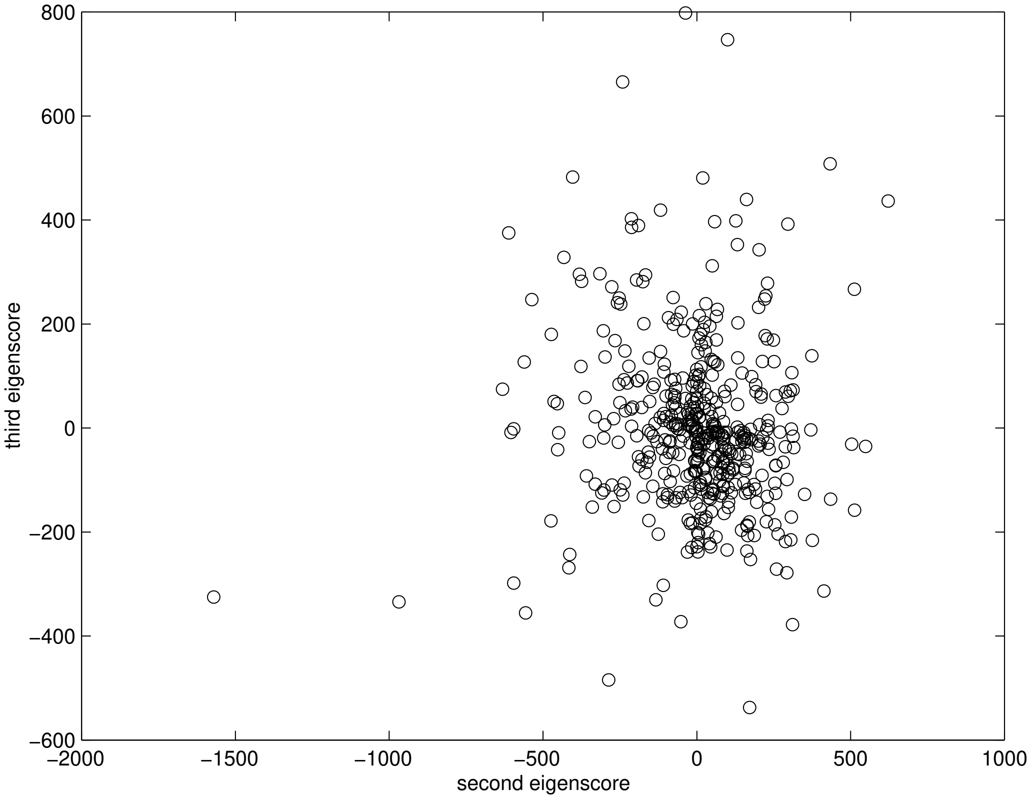

Figure 9: Scores on second and third eigenvectors

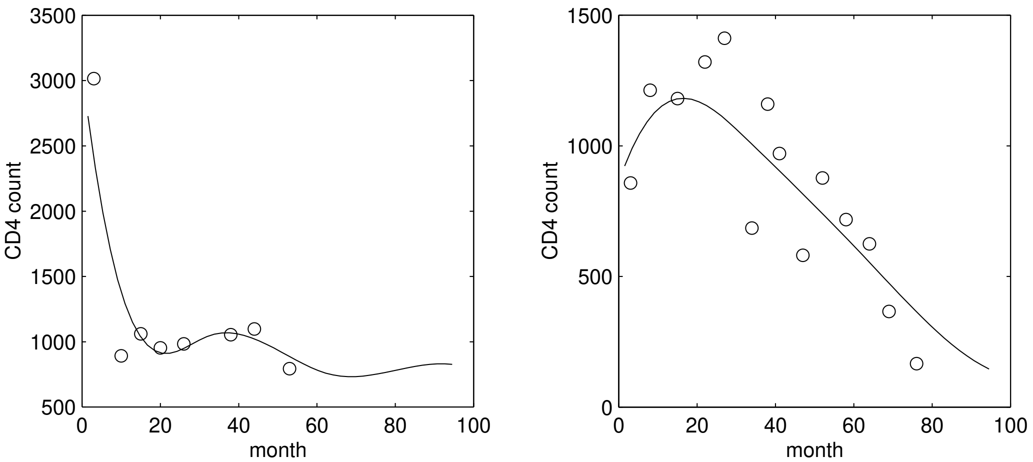

Figure 10: CD4 counts and BLUP smooths for cases with the lowest score on the second eigenvector (left) and the highest score on the third eigenvector (right).

100 resulting estimates). The covariate effect is small, and even its sign cannot be reliably determined. Smoking, on the other hand, is associated with a increased, but possibly declining, level of CD4. As noted in Zeger and Diggle (1994), this may be due to healthier men continuing to smoke.

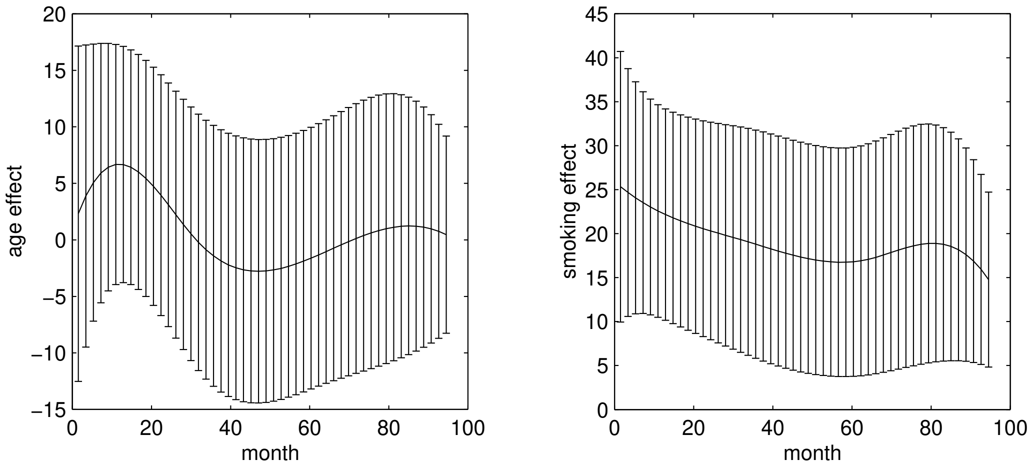

Figure 11: The "effect" curves and one standard deviation error bars found by a bootstrap for age (left) and smoking status (right).

4 Concluding Remarks

We have provided a simple yet flexible and powerful extension of classical linear mixed effects modeling to explicitly non-parametric models. By using a basis function approach, the classical conceptual and computational methodology transfers directly to a non-parametric setting. The population mean function can be estimated as well as smoothed individual trajectories. A flexible framework for estimating association of covariates with changes in curve shapes is provided.

We mention a number of points in closing. First, we note that bases other than splines could clearly be used in this application of the method of sieves (Grenander, 1981). One of the advantages of using a low dimensional approximation rather than penalyzing a high dimensional approximation or using kernel smoothing is computational—we do not have to solve large linear systems. The analysis of the CD4 counts reported in the previous section, including the bootstraps, coded in Matlab, took about three minutes on a Sun Ultra 2. Model selection criteria are a key aspect of our methodology; they are seemingly effective, but a clearer understanding of their properties in this context would be valuable. Our approach to selecting a spline basis has been fairly crude—we have only allowed equally spaced breakpoints. It might be possible to obtain finer resolution by attempting to optimize breakpoint locations as in Stone et al. (1997), but at least for exploratory work such as in our examples, it is not clear that gain would offset the additional computational effort.

In our analyses we have explicitly decomposed the random trajectories into a common mean function of time and random deviations from that mean, but we would like to point out that it is not really necessary to do so. If each random trajectory is simply modeled as a linear combination of basis functions with random coefficients, the mean function is specified by the expected values of those coefficients.

We have limited our attention to time independent covariates, but in principle the methodology can be extended to time varying covariates. For example, we might wish to predict the future course of a trajectory based on observation of it up to the present. We plan to pursue this direction in future research.

We hope that the methodology we have presented will be useful in data mining large collections of irregularly sampled random curves. By summarizing them as BLUP estimates of coefficients with respect to some basis, standard multivariate methods become applicable. Methods of outlier identification and clustering can be applied, for example, or the coefficients could become inputs for classification trees.

References

Besse, P., Cardot, H. and Ferraty, F. (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data, *Computational Statistics and Data Analysis* **24**: 255–270.

Brumback, B. and Rice, J. (1997). Smoothing spline models for the analysis of nested and crossed samples of curves, *Journal of the American Statistical Association* . (to appear).

Diggle, P., Liang, K.-Y. and Zeger, S. (1994). *Analysis of Longitudinal Data*, Oxford Science Publications.

Fan, J. and Zhang, J.-T. (1997). Functional linear models for longitudinal data. manuscript.

Gasser, T., Müller, H.-G., Köhler, W., Molinari, L. and Prader, A. (1984). Nonparametric regression analysis of growth curves, *Annals of Statistics* **12**: 210–229.

Grenander, U. (1981). *Abstract Inference*, Wiley.

Hoover, D., Rice, J., Wu, C. and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika* . (to appear).

Jones, R. (1993). *Longitudinal Data with Serial Correlation: a State Space Approach*, Chapman and Hall.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. and Rinaldo, C. R. (1987). The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of participants, *American Journal of Epidemiology* **126**: 310–318.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data, *Biometrics* **38**: 963–974.

Olshen, R., Biden, E., Wyatt, M. and Sutherland, D. (1989). Gait analysis and the bootstrap, *Annals of Statistics* **17**: 1419–1440.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*, Springer.

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

Robinson, G. (1991). That BLUP is a good thing: The estimation of random effects, *Statistical Science* **6**: 15–32.

Stone, C., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling, *Annals of Statistics* **25**.

Vonesh, E. and Chincilli, V. (1977). *Linear and nonlinear models for the analysis of repeated measurements*, Marcel Dekker.

Zeger, S. and Diggle, P. (1994). Semi-parametric models for longitudinal data with applications to CD4 cell numbers in HIV seroconverters, *Biometrics* **50**: 689–699.