BIOMETRICS 57, 253-259
March 2001

Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves

John A. Rice

Department of Statistics, University of California at Berkeley,
Berkeley, California 94720, U.S.A.
email: rice@stat.berkeley.edu

and

Colin O. Wu

Department of Mathematical Sciences, The Johns Hopkins University,
Baltimore, Maryland 21218, U.S.A.

SUMMARY. 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 best linear unbiased predictor (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—Akaike information criterion (AIC), Bayesian information criterion (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.

KEY WORDS: Functional data analysis; Longitudinal data; Spline.

1. Introduction

In recent years, there has been an increasing interest in nonparametric 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). In this article, we present a methodology for analysis of collections of curves that may be unequally and sparsely sampled, in contrast with the usual requirement that the observations be equally spaced or nearly so. We propose a nonparametric methodology for the following purposes: (1) estimation of the population mean curve, (2) estimation of the covariance function without a requirement that it be stationary or of any particular parametric form, (3) estimation of the eigenfunctions and eigenvalues of the covariance function, (4) smoothing individual curves, (5) estimation of functionals of individual curves, such as derivatives and location of extrema, (6) estimation of the effects of covariates on the shapes of the curves, and (7) exploration of the patterns of variability among the individual curves and identification of unusual ones. In achieving these goals, computational efficiency is important.

To accomplish these aims, we extend current methodology for analyzing repeated measures data via linear mixed effects models to a nonparametric setting. A typical parametric mixed effects analysis of this type 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. Our extension consists of approximating 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 point of view is nonparametric (the method of sieves; Grenander, 1981), with the dimensionality of the spline approximation adaptively chosen via model selection techniques.

Eigenfunctions of the estimated covariance functions can be easily computed and provide insight into the modes of variability present among the individual curves. Smooth curves (spline functions) for individual trajectories are constructed as BLUP estimates (Robinson, 1991), combining data from that individual and the entire collection. These smoothers may be viewed as kernels whose local shape is adaptively tailored to the particular collection of curves to be smoothed. The influence of covariates on the shapes of curves is naturally quantified by examining the dependence of the spline coefficients upon them.

Currently existing methods, although related, do not accomplish all the aims set out in the first paragraph above. The method is related to that of Brumback and Rice (1998), who also use splines but assume a particular form of the covariance kernel arising from smoothing splines. 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.) In Rice and Silverman (1991), the data are assumed to be collected on a regular grid. Fan and Zhang (unpublished manuscript) and Hoover et al. (1998) treat nonparametric estimation of the mean function but not the covariance function. Diggle and Verbyla (1998) construct an estimate of the covariance function of repeated measures data by locally smoothing empirical variograms but not a direct estimate of random effects. In contrast, our approach produces an estimate of the covariance function by approximating the random effects and provides a structure for modeling covariate effects. Staniswalis and Lee (1998) approximate the curves by eigenfunctions derived from a covariance function estimated by kernel smoothing. Our procedures use splines rather than kernels for nonparametric estimation and adaptively select the numbers of knots.

We present two real examples and a simulation. The first example is a collection of the curves formed by the angles of the hip over walking cycles of 39 children (Olshen et al., 1989). 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 that seemed reasonable, especially in light of the large signal-to-noise ratio. Here we do not interpolate but use our nonparametric random effects methodology to estimate the covariance function and its eigenfunctions directly from the original measurements. We also show how information from the entire collection of curves can be used to smooth and differentiate individual trajectories (see Ramsay and Silverman (1997) for discussion of the importance of derivatives in modeling functional data).

As a second example, we consider sequences of CD4 counts from 463 homosexual men from the Multi-Center AIDS Cohort Study (Kaslow et al., 1987) who seroconverted between 1984 and 1993. 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 with the gait data, the observations are noisy and sparse and, other than a decreasing trend, dynamic 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 are time independent). Since there is no natural *a priori* parametric model for the mean and covariance function of these data, a nonparametric approach is desirable. We show how our nonparametric estimate of the covariance function provides a summary of the time dependence in the individual trajectories and how its eigenfunctions can inform understanding and exploration of their modes of variability. Our nonparametric methodology reveals no relation between age at onset of seroconversion and the shape of the trajectory and a weak relationship of the shape to smoking status.

The remainder of the article is organized as follows. In Section 2, we present the general methodology. Section 3 is devoted to the two real examples. Section 4 reports the results of a simulation, and Section 5 contains some closing comments.

2. Methodology

Let there be m subjects, n_i observations at times $0 \le t_{ij} \le T$ on the ith subject, and $n = \Sigma_{i=1}^{m} n_i$ observations in all. Let $Y_{ij} = Y_i(t_{ij})$ be the outcome measured on the ith 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 approximate the mean function and the random function nonparametrically with splines. The mean function is

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

where $\{\bar{B}_k(\cdot)\}$ 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 ith subject is similarly modeled as the spline function $\Sigma_{k=1}^{q} \gamma_{ik} B_k(t)$. Here $\{B_k(\cdot)\}$ is a basis for a possibly different space of spline functions on $[0, T]$ and the γ_{ik} are random coefficients with mean zero and covariance matrix Γ. Finally, incorporating uncorrelated measurement errors ϵ_{ij} with mean zero and constant variance σ^2, our approximating model is

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

The covariance structure, including serial correlation, is modeled through the γ_{ik}. The covariance kernel for a random curve $Y(t)$ is thus approximated as

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

where $\delta(\cdot)$ 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.

Before continuing, we make some further remarks on the nature of this model. We view the sample path for each individual as composed of a smooth random function, not necessarily stationary, measured at discrete time points with white measurement error. If there were multiple measurements for each individual, a further decomposition would be needed to model within-subject variation, but since we only deal here with a single realization for each individual, the components of such a decomposition are unidentifiable. If the measurement process were such as to induce serial correlation among the errors, our model would not account for that properly. Finally, we note again that our structure (2) is a sieve approximation to this underlying model. Conditioning on p and q, (2) is a classical linear mixed effects model, and the vector of observations on the ith 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 $V_i = Z_i\Gamma Z_i^{\mathrm{T}} + \sigma^2 I$. 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

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

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. The estimate thus uses all the curves to smooth a single one by using information in the collection to construct a kernel whose shape and bandwidth vary locally and by using the mean function of the collection. 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. Note that derivatives are easily calculated from the representation (6).

There is no 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 alternative form of eigen analysis based on the eigenvectors of Γ seems less desirable because it depends on the particular basis used for the spline functions.) 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.

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. 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, i.e.,

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

Here $\mu_i = (\mu(t_{i1}), \ldots, \mu(t_{in_i}))^{\mathrm{T}}$. In the examples that follow, we also use AIC and BIC, which give results similar to those obtained by cross-validation and are faster to compute. In our nonparametric context, procedures directly focused on model selection seem more appropriate than hypothesis testing procedures, such as in Lin (1997).

We now discuss the incorporation of covariates. First, linear fixed effects can be included by adding columns to the design matrices X_i. 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),

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

The change in curve shape due to covariates is thus modeled through the change in the coefficients. We refer to the second term in (8) as the effect curve of the covariates.

For a varying-coefficients linear model,

$$\mathrm{E}(\gamma \mid U = u) = \Sigma_{\gamma U}\Sigma_{UU}^{-1}(u - \mathrm{E}(U)), \tag{9}$$

$$\mathrm{E}(\epsilon \mid U = u) = \Sigma_{\epsilon U}\Sigma_{UU}^{-1}(u - \mathrm{E}(U)). \tag{10}$$

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 article, we only consider such linear estimates for $\mathrm{E}(\gamma_{ik} \mid U = u)$, but the possibility of using more sophisticated methods (such as nonparametric regression) for predicting the random spline coefficients by covariates is evident.

3. Examples

3.1 *Human Gait*

The data are described in the Introduction. Figure 1 shows the cross-validated log likelihood, the AIC criterion, and the BIC

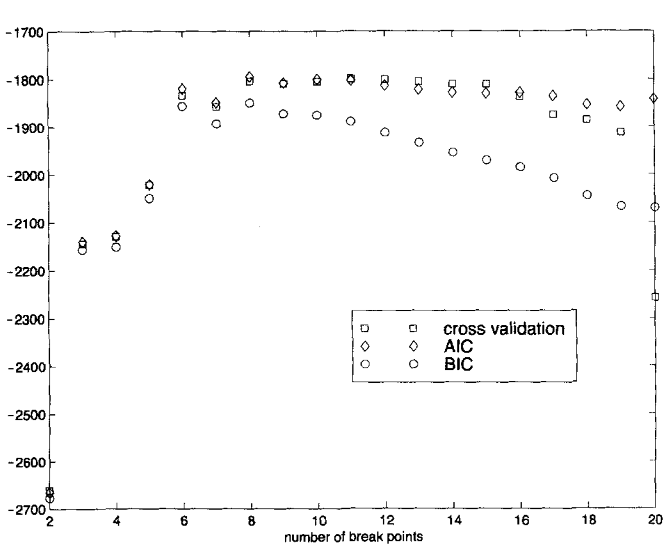

Figure 1. Model selection criteria for the number of breakpoints for the hip-angle curves.

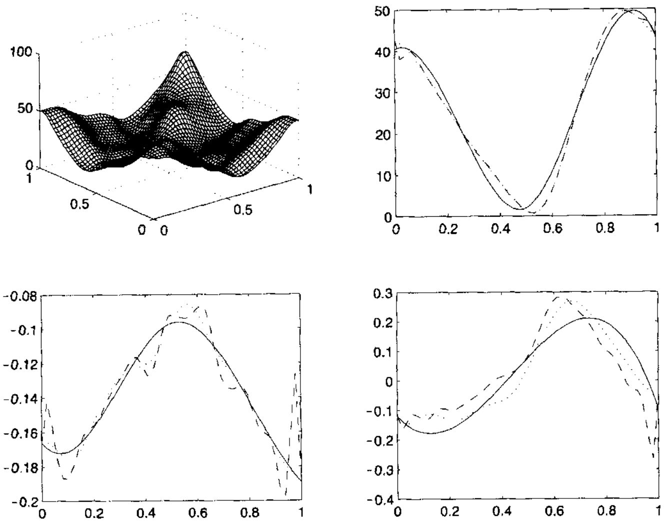

Figure 2 Top left: Covariance function for a hip cycle estimated with 10 breakpoints. Top right: Estimates of mean function with 3 breakpoints (solid), 10 breakpoints (dotted), and 20 breakpoints (dashed). Bottom left: Estimates of the first eigenfunction. Bottom right: Estimates of the second eigenfunction.

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 zero and one, so a pure cubic has two breakpoints, for example. The cross-validation function and the AIC criterion are rather flat in the region from 6 to about 15 breakpoints, whereas the BIC criterion drops more rapidly after reaching a maximum at 8 breakpoints. The high signal-to-noise ratio supports a fairly high-dimensional approximation (we also experimented by adding noise to the data, and indeed the criteria then peaked at lower dimensions).

The covariance function estimated from 10 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 figure also shows the estimated mean and the first two eigenfunctions of the covariance function for 3, 10, and 20 breakpoints. Ten and 20 breakpoints give very similar estimates of the mean curve, whereas there is apparently noticeable systematic distortion when 3 are used. For the covariance structure, the major effect of overfitting is roughness near the boundaries. The results for 10 breakpoints are very similar to those obtained in Rice and Silverman (1991). For interpretation of the eigenfunctions, we refer the reader to that paper.

The left panel of Figure 3 shows the BLUP smoothing (6) of a single curve for 3, 10, and 20 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 2 and discussed above. The choice of three breakpoints oversmooths the data. From the 10 breakpoint smoothing, the curve can be interpolated at times between the observation points and the ordinates and abscissa of extrema can be found. First and second derivatives can be easily calculated using the B-spline coefficients.

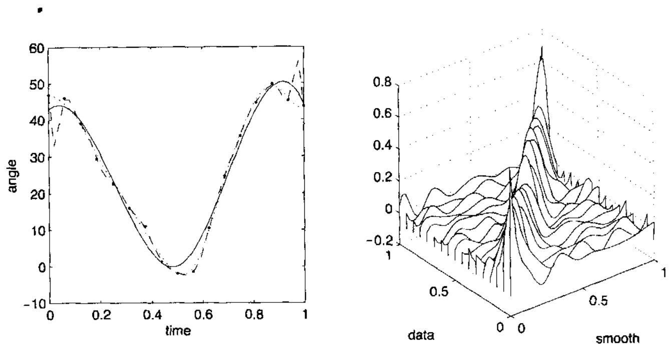

Figure 3. Left: Three smoothings of the data from a single individual with 3 breakpoints (solid), 10 breakpoints (dotted), and 20 breakpoints (dashed). Right: The smoother matrix for 10 breakpoints; lines show the weights of each data point as a function of the time at which the smoothing is computed.

The right panel of Figure 3 displays the hat matrix, or kernel, which maps observed data points into the smoothed values shown in the left panel for 10 breakpoints. For this subject, the nonparametric smoother has fairly constant shape and bandwidth, supported on an interval of width about .2 cycles, with some apparent modification of the weighting kernel near the boundaries.

3.2 CD4 *Counts*

Similar data were analyzed by Zeger and Diggle (1994) using a semiparametric model and by Fan and Zhang (unpublished manuscript) to illustrate the use of a particular functional linear model. We first consider modeling the evolution of the CD4 counts with time only. 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 4 along with the individual trajectories. The cova-

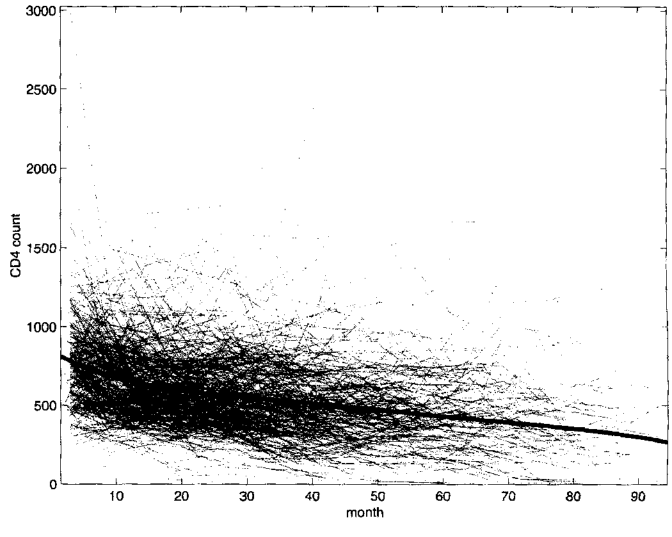

Figure 4. Estimated mean function and individual trajectories of 463 sequences of CD4 counts.

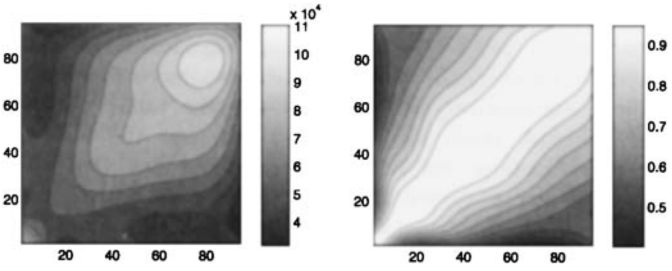

Figure 5. Left panel: Contours of the estimated covariance function of CD4 counts. Right panel: Corresonding correlation function.

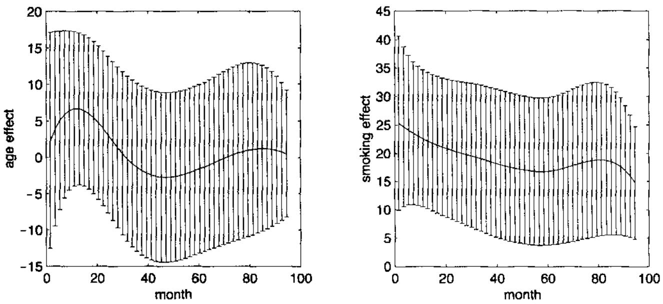

Figure 7. The effect curves and 1 SD error bars found by a bootstrap for age (left) and smoking status (right).

riance function is shown in Figure 5; it is clearly nonstationary with high variability at very early times, decreasing until about 20 days and then increasing. Variability at early times may be in part due to lack of precision in identifying the actual date of seroconversion. The dependence in the curves can also be quantified by the correlation function, shown in Figure 5. Here it is seen that there is quite strong correlation between all times. However, the correlation between very early counts and later counts dies off relatively rapidly, whereas for middle and later times, the dependence persists more strongly. Early counts thus have relatively less predictive value. These sorts of features would be difficult to anticipate in a traditional parametric model; they would not be produced, e.g., by linear random effects model with only a random intercept and slope.

We next consider the eigenfunction decomposition of the estimated covariance function. The first eigenfunction corresponds to an overall shift of CD4 level, the second to a trend, the third to a rapid change in later months, and the fourth (if the coefficient is positive) to an increase in the early and late months and a dip in the middle months. These eigenfunctions, respectively, account for 83, 10, 5, and 2% of the total variance. Thus, by far, most of the variability is accounted for by shifts in overall level. The first two eigenfunctions are shown in Figure 6.

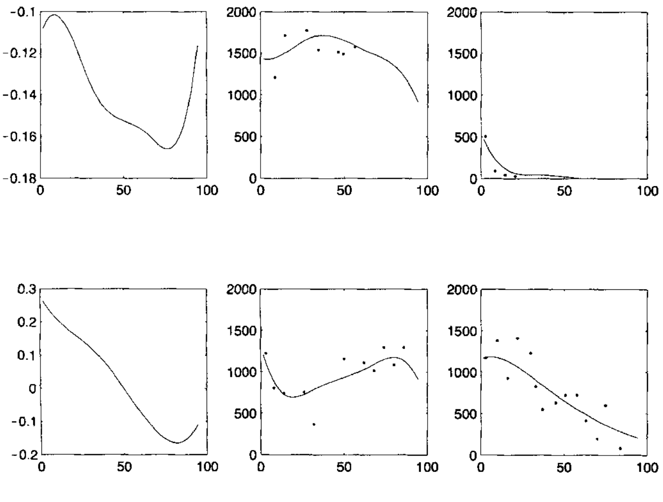

Figure 6. The first two eigenfunctions of the estimated CD4 count covariance function are shown in the left column. The observations and smoothed curves of extreme cases in those directions are shown in the corresponding center and right columns.

In exploring such a data set, 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. One way to begin to visually explore the variability in the set is to single out cases with extreme projections in the directions of the eigenfunctions, the projections being found using the inner products of the BLUP smooths and the eigenfunctions. Four such cases, corresponding to largest positive and largest negative projections on the first two eigenfunctions, are shown in Figure 6. By this technique, some curves that differ in interesting ways from the mean curve are extracted from the blur of Figure 4. For example, cases with unusually strong positive and negative trends are shown in the second row of the figure.

We examined the effects of age and smoking status. We standardized subjects' ages, which ranged from 18 to 64. Figure 7 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 (conditional on four breakpoints) found by the bootstrap as in Hoover et al. (1998) (subjects were sampled with replacement 100 times; the error bars are the pointwise standard deviations of the 100 resulting estimates). The covariate effect is small—even its sign cannot be reliably determined. Smoking, on the other hand, is associated with an increased, but possibly declining, level of CD4. As noted in Zeger and Diggle (1994), this may be due to healthier men continuing to smoke.

4. A Synthetic Example

In order to gauge the performance of our methodology, we conducted a small simulation. The covariance function of the process was constructed from Legendre polynomials on $[-1, 1]$, multiplied by $\exp(-x^2/6)$ in order that the structure of the covariance function not be purely polynomial. A single sample was made up of 100 partially observed paths. Each was observed at points, the number of which was chosen from a discrete uniform distribution on $[1, \ldots, 10]$ and the locations of which were uniform on $[-1, 1]$. At each of those points, the observation consisted of the value of the sample path plus random Gaussian error with mean zero and standard

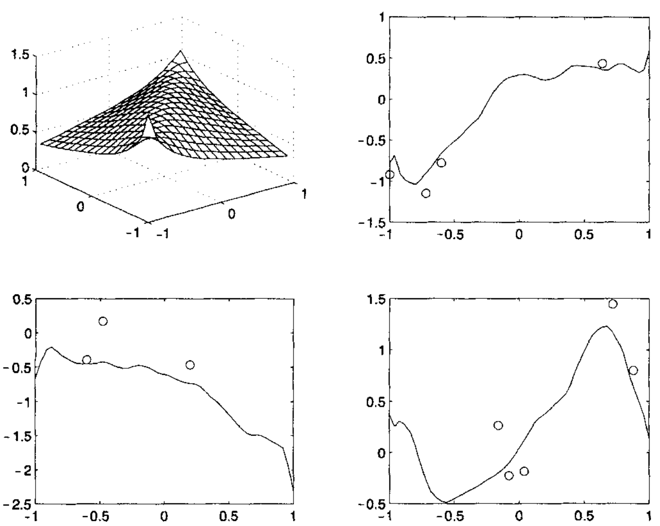

Figure 8. The covariance function, three random trajectories, and observations corrupted by noise with SD .25.

deviation .25. Figure 8 shows the covariance function, three random trajectories, and the resulting observations. There were 100 such samples in all, and from each the mean and covariance function were estimated as described above, with equally spaced break points chosen by BIC.

It is not entirely obvious how to meaningfully evaluate the estimates, since evaluation depends on the uses to which the estimates are to be put. We considered two summary measures. The first uses the estimated covariance function to form the BLUP estimates at the observation points that occurred in the sample. We judge the result with respect to the best estimate, which uses the true covariance function. Pooling over all points and all runs, using the estimated covariance functions resulted in an average increase in root mean squared prediction error 5% greater than that of the optimal estimate. The standard deviation between runs was 2.3%. For contrast, using just the observed values resulted in an increase of 60% with a standard deviation of 10%. Thus, from the standpoint of smoothing the observed responses, the estimated covariance functions were quite effective.

A second way of evaluating the estimated covariance function is to consider the shapes of the sample paths it would produce. The sample paths produced by the true covariance function are linear combinations of its eigenfunctions with root mean squared amplitudes given by the square roots of the corresponding eigenvalues. We thus examined the root mean square projections onto those eigenfunctions that would be produced by the estimated covariance functions. Table 1 summarizes the results. From the table, we see that the root mean square projections onto the leading five eigenfunctions are accurately produced but that the higher frequency contributions to the sample paths from the remaining eigenfunctions are dampened. In summary, the estimated covariance functions correspond to sample paths with dominant low-frequency features resembling those produced by the actual covariance function but with less high-frequency content.

5. Concluding Remarks

We mention a number of points in closing. First, we note that bases other than splines could clearly be used. One of the advantages of using a low-dimensional approximation rather than penalizing a high-dimensional approximation or using kernel smoothing is computational—we do not have to solve large linear systems. Model selection criteria are a key aspect of our methodology; they are seemingly effective, but a clearer understanding of their empirical and theoretical 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. The work of Stone et al. (1997) and Smith and Kohn (1996) suggests possibilities for variable knot selection in our multiple-curve problem, but since their methods only deal with a single curve, extending their work appropriately would require substantial further development.

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. In such an approach, the basis functions are the same for the mean and the random effects, which we have found to be adequate in our examples.

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

Table 1

The first row shows the root mean square (RMS) projections on the first 10 *eigenfunctions as determined from the actual covariance function. The second row shows the mean RMS projections produced by the estimated covariance functions in the simulation. The last row shows the standard deviations over the* 100 *runs of the simulation corresponding to the entries in the second row.*

Actual	6.2	2.7	1.8	1.2	.81	.58	.43	.33	.25	.20
Mean of estimated	6.1	2.7	1.8	1.2	.84	.10	.04	.01	.01	.01
SD of estimated	.51	.29	.25	.25	.18	.02	.01	.003	.002	.002

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, e.g., or the coefficients could become inputs for classification trees.

Acknowledgements

Partial support was provided by the National Institute on Drug Abuse grant RO1 DA10184-01. Additional support was provided by the National Science Foundation (DMS-9625384) for the first author and by the A. Duncan Fund of the Johns Hopkins University for the second author. The authors would like to thank Prof Joseph Margolick (Johns Hopkins) for providing the MACS Public Use Data Set Release PO4 (1984–1991), and the reviewers for helpful comments and suggestions.

Résumé

Nous proposons une méthode pour analyser des familles de courbes telles que le modèle de courbe individuelle soit composé de fonctions splines à coefficients aléatoires. La méthode s'applique lorsque les courbes individuelles sont échantillonnées même à des points variables et irrégulièrement espacés. On obtient une approximation de basse fréquence pour la structure de covariance, qui peut être estimée par l'algorithme EM. Des courbes lissées pour les trajectoires individuelles sont obtenues comme estimateurs BLUP, en combinant les données de l'individu et celles de toute la famille. Ce cadre conduit naturellement à des méthodes pour examiner l'effet de covariables sur la forme des courbes. Nous utilisons des techniques de sélection de modèles—AIC, BIC et validation croisée—pour choisir le nombre de points de rupture pour l'approximation spline. Cette méthodologie fournit des outils simples, flexibles et efficaces du point de vue du calcul, pour l'analyse de données fonctionnelles.

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. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93,** 944–961.

Diggle, P. J. and Verbyla, A. P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* **54,** 401–415.

Grenander, U. (1981). *Abstract Inference.* New York: Wiley.

Hoover, D., Rice, J., Wu, C., and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85,** 809–822.

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.

Lin, X. (1997). Variance component testing in generalized linear models with random effects. *Biometrika* **84,** 317–343.

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.* New York: 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.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75,** 317–343.

Staniswalis, J. and Lee, J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **93,** 1403–1418.

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,** 1371–1425.

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.

Received June 1998. *Revised July* 2000.
Accepted July 2000.