PSYCHOMETRIKA—VOL. 51, NO. 2, 285–311
JUNE 1986

PRINCIPAL COMPONENTS ANALYSIS OF SAMPLED FUNCTIONS

PHILIPPE BESSE

UNIVERSITE PAUL SABATIER, TOULOUSE, FRANCE

J. O. RAMSAY

MCGILL UNIVERSITY

This paper describes a technique for principal components analysis of data consisting of n functions each observed at p argument values. This problem arises particularly in the analysis of longitudinal data in which some behavior of a number of subjects is measured at a number of points in time. In such cases information about the behavior of one or more derivatives of the function being sampled can often be very useful, as for example in the analysis of growth or learning curves. It is shown that the use of derivative information is equivalent to a change of metric for the row space in classical principal components analysis. The reproducing kernel for the Hilbert space of functions plays a central role, and defines the best interpolating functions, which are generalized spline functions. An example is offered of how sensitivity to derivative information can reveal interesting aspects of the data.

Key words: reproducing kernel, Hilbert space of functions, spline functions, Green's functions, interpolation, smoothing.

1. Introduction

This paper considers data which are replications of sampled functions. That is, there is a set of p distinct points t_j and a set of n functions with values $x(t)$ yielding the data $x_i(t_j)$, $i = 1, \ldots, n; j = 1, \ldots, p$. In the behavioral sciences longitudinal data are typically of this type; each of n persons is measured on some univariate variable at times t_1, t_2, and so on. Curves of learning and forgetting, scores for examinees tested repeatedly, subjective or physiological responses over time, and dose response functions are typical examples. A set of histograms can be regarded as a set of density functions observed as a finite number of points. A set of periodograms arising from classical time series analysis can be viewed as a continuous spectrum sampled at discrete points in the frequency domain. A set of psychophysical functions giving a subjective value corresponding to each of a finite number of physical magnitudes for each of a number of observers is yet another example.

Models for longitudinal data of this type often include the hypothesis that at least part of the variation of the data can be accounted for in terms of linear combinations of known functions. That is, one can propose the decomposition

$$x_i(t) = \sum_{k=1}^{m} c_{ik} u_k(t) + e_i(t), \tag{1}$$

where the known functions u_k represent the predicted part of the sampled function $x_i(t)$

This research was supported by Grant PA 0320 to the second author by the Natural Sciences and Research Council of Canada. We are grateful to the reviewers of an earlier version and to J. B. Kruskal and S. Winsberg for many helpful comments concerning exposition.

Requests for reprints should be sent to J. O. Ramsay, Department of Psychology, 1205 Dr. Penfield Ave., Montréal, Québec, H3A 1B1, Canada.

0033-3123/86/0600-6081$00.75/0

and the remainder $e_i(t)$ represents that residual variation which is to be studied further or regarded as noise.

Data of this sort are often analyzed in one of two ways. The first approach is to regard them as a candidate for classical time series analysis after, possibly, removing any systematic trends by regression analysis (Anderson, 1971). Here the rationale is usually that removal of the components u_k from x_i will leave a residual that is a stationary autoregressive and/or moving average process which is regarded as the noise component. If the trend in the data is linear, it is usual to analyze first differences in x rather than the original sampled values. Thus, time series analysis as usually applied makes rather strong assumptions such as stationarity about the residual covariance structure in the data that may not always be plausible. A second approach (Rao, 1958; Tucker, 1958) has been to regard the data as n p-variate observations and to employ multivariate techniques such as principal components analysis (PCA) or factor analysis. Here the known components u_k can be regarded as instrumental variables which may or may not be incorporated into the analysis (Rao, 1964).

Both the time series and PCA approaches may be summarized as follows: let $\|\cdot\|$ be a norm defined on the vector space of symmetric real matrices of order p. Then the objective is

$$\min_{A,B,C} \| X - CU - AB \|, \tag{2}$$

where U is the m by p matrix of values $u_k(t_j)$, and A, B, and C are to be estimated and are n by r, r by p, and n by m, respectively, with $r \leq p - m$. Rao (1980) discusses this general problem in the context of norms of the form

$$\| Y \|^2 = \mathrm{tr}(Y'WYM). \tag{3}$$

The symmetric positive definite matrices W and M of orders n and p, respectively, define the metric within which the analysis is carried out. Since it will be assumed here that the replicates are independent, it is reasonable to assume $W = I$, but this leaves open the question of choice of M. Within the context of time series analysis, the Gauss-Markov theorem tells us that this should be an estimate of Σ^{-1}, where Σ is the population covariance of the residuals resulting from the regression of a sampled function x_i on the u_k's. However, in practice this matrix will be unknown. In any case, where it is known that sampled values do not have any appreciable error of observation, the interpretation of the Gauss-Markov theorem is not obvious. In PCA applications, on the other hand, it is usual to use $M = I$, although more generally PCA can be carried out in any metric. Thus, both time series analysis and PCA raise the problem of how to choose the metric for the analysis.

It is the purpose of this paper to motivate the choice of metric by taking a functional analytic view of the regression and principal components analysis of sampled functions. By this we mean that we shall consider the data as arising from the observation of a set of n random functions at discrete points in time. These random functions will be presumed to have a certain level of smoothness, and we hope to show how this assumption can affect the data analysis. In particular, the smoothness assumptions plus statements about how the function space within which they lie can be partitioned will lead to a settling of the metric question in the classical approaches. Thus, our approach does not so much offer a competitor to classical approaches as it complements them by determining the metric for the least squares analysis. Broadly speaking, the two goals in this approach are to display a theoretical rationale for traditional approaches, and to provide a family of practical procedures which have the potential of revealing new and interesting aspects of the data.

Our approach will draw on three areas that do not often appear in the pages of applied statistics journals: the properties of generalized spline functions, solutions to ordinary and boundary value linear differential equations, and the theory of reproducing kernel Hilbert spaces. We will not attempt to state and prove all the needed results; rather our aim is to provide an account of how these topics relate to applied data analysis and perhaps provoke an interest in further reading. As a consequence, the treatment will be kept relatively informal. A more formal presentation is in preparation, and will appear elsewhere.

2. The Tongue Data

In order to provide a concrete problem which illustrates our approach, we will discuss the data presented in Table 1. The production of speech involves the movement of various critical parts of the vocal tract. The back of the tongue and the soft palate is one such articulation system, and it is implicated in the production of vowel sounds, the stops "*k*" and "*g*," and their associated fricatives. A central question is how such a system is controlled by the central nervous system during the exceedingly rapid and complex movements required by speech.

The data consist of 42 records of tongue dorsum height collected by Munhall (1984) using an ultrasound sensing technique developed by Keller and Ostry (1983). Each record arose from the utterance of the sound "kah" at the beginning of which the tongue was in contact with the soft palate. A good deal of preprocessing involving cubic spline smoothing of observations at every millisecond was preliminary to the data in Table 1, which can be considered to begin and end at points where the tongue height had zero velocity and to have negligible error or noise components. It will be assumed that the interval of observation has been normalized to be $[0, \pi]$. Thus, the sampled values in Table 1 can be reasonably regarded as errorless observations of 42 cubic spline functions at 13 equally spaced points. These are displayed in Figure 1.

Inspection of these curves reveals that most of them can be fairly well summarized by the model

$$x_i(t) = c_{i1} + c_{i2} \sin (t) + c_{i3} \cos (t) + e_i(t). \tag{4}$$

This model is consistent with the hypothesis that the tongue is acting like a spring which has been set in motion and left free of oscillate, and for which the damping factor is too small to be noticed within a single oscillation. Thus, to a first order of approximation, tongue motion appears to be obeying the linear differential equation

$$Dx + D^3x = 0, \tag{5}$$

since any linear combination of the constant, sine, and cosine functions would satisfy such an equation. It is indeed the case that muscle tissue has strong spring-like characteristics (Hunter & Kearney, 1982; Huxley, 1980), where the damping factor is fairly small and where the period of oscillation is controlled by the tension between opposing muscle groups. The interesting question here is to what extent there is interesting variation in tongue dorsum behavior beyond these well-known components. Put another way, how will any input from the central nervous system manifest itself if at all? Can the tongue be regarded as a system in ballistic motion free of outside control, or will it be clear that its motion is affected by some form of external input?

3. Ordinary Least Squares Analysis of the Tongue Data

Figure 2 displays the sampled values of the variance function computed by com-

Table 1.
Smoothed Records of Tongue Dorsum Displacement
Measured by Ultra-sound Sensing at
13 Equally Spaced Time Points (Units Arbitrary)

	Sampling Point												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	587	585	579	570	558	545	531	519	509	502	497	494	493
2	591	590	584	576	565	553	541	529	520	512	507	505	504
3	588	587	584	578	571	563	553	543	534	527	523	521	520
4	587	584	575	562	547	533	522	515	511	508	505	503	502
5	585	584	579	571	560	548	537	526	517	512	509	508	508
6	585	583	577	567	556	544	532	522	514	508	503	499	498
7	616	613	605	595	584	574	565	558	552	548	545	543	543
8	623	621	615	604	590	575	562	550	541	535	531	529	529
9	630	626	616	604	591	577	563	550	540	534	531	529	529
10	628	626	620	612	601	588	573	560	548	540	536	533	533
11	618	615	608	596	582	568	554	544	536	531	527	525	525
12	596	595	591	583	574	564	555	547	540	536	533	531	531
13	611	609	604	595	583	570	557	545	536	531	528	526	526
14	612	609	603	593	582	569	558	548	541	535	532	530	530
15	609	607	603	596	588	578	568	559	551	545	541	539	538
16	608	606	600	590	578	565	554	545	539	534	531	529	529
17	601	600	597	593	587	578	569	560	553	547	543	540	539
18	571	568	561	551	540	528	518	510	504	499	494	491	490
19	556	556	554	549	541	531	521	511	503	496	491	488	488
20	572	571	567	561	552	539	526	513	503	495	489	486	486
21	578	577	574	568	561	553	544	536	528	522	518	516	515
22	588	587	582	573	561	550	539	531	525	522	522	521	521
23	576	574	568	558	546	535	525	517	511	508	505	503	502
24	579	578	573	565	554	543	532	522	514	507	502	500	499
25	569	568	563	557	549	541	531	522	513	506	502	500	499
26	582	581	577	570	560	549	538	529	520	514	510	508	507
27	585	584	579	572	563	553	543	534	526	519	515	513	512
28	582	581	577	570	560	549	538	529	520	514	510	508	507
29	589	587	583	576	566	554	543	532	524	517	513	511	511
30	601	600	595	586	574	561	548	538	529	524	520	519	518
31	590	589	585	578	570	560	550	541	532	525	521	518	517
32	590	588	583	576	568	558	547	536	528	522	517	515	515
33	587	585	582	575	567	556	545	535	527	522	519	517	516
34	601	599	595	587	577	565	553	542	533	527	523	521	520
35	576	575	572	569	564	558	550	540	531	523	518	516	516
36	588	586	580	572	560	547	535	525	518	515	512	511	510
37	589	587	582	572	561	548	536	526	519	515	513	511	511
38	585	583	579	572	564	555	545	537	530	525	522	520	520
39	585	583	579	572	563	554	546	539	534	531	528	526	525
40	584	583	579	572	564	555	547	540	534	529	526	524	524
41	587	585	579	572	563	553	542	533	527	523	521	520	519
42	584	583	580	574	567	559	551	544	539	536	534	533	533

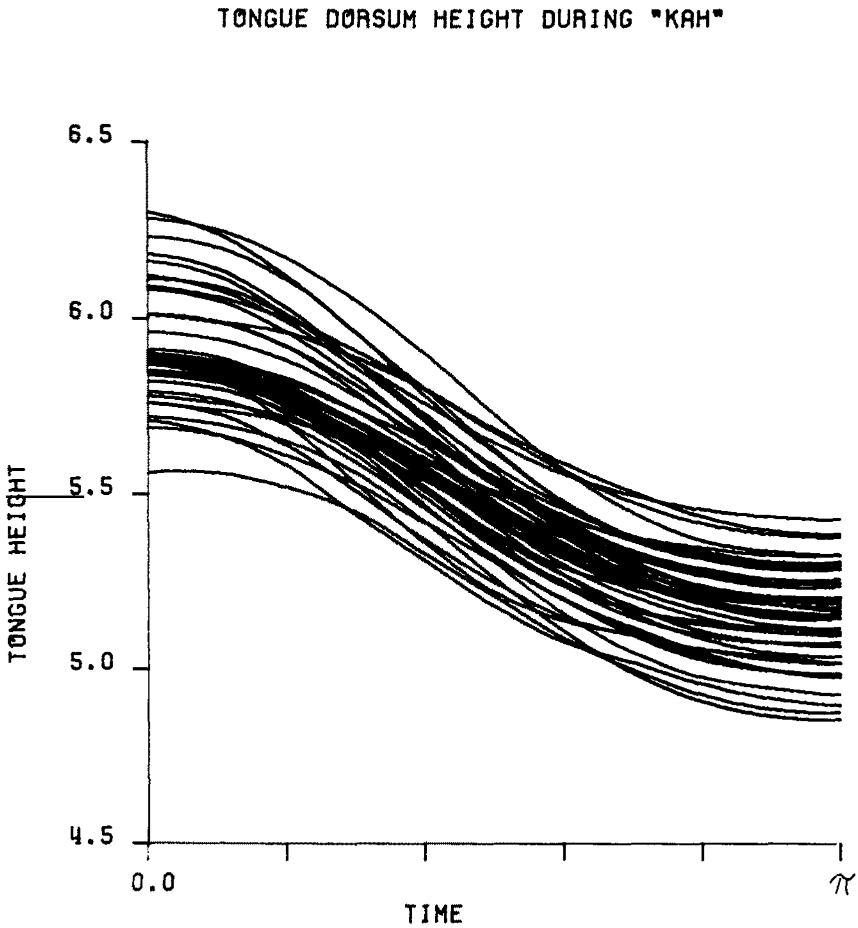

FIGURE 1

The movement of the back of the tongue (units arbitrary) during the utterance of the sound "kah." The curves are a result of polynomial spline smoothing of tongue position sampled every millisecond using an ultrasound sensing technique. Each record begins and ends at the point where the slope is zero, and the lengths of the curves have been standardized to the interval $[0, \pi]$.

puting the variance of the sampled values at each of the 13 sampling points, as well as corresponding variance function values. The mean function confirms at a visual level, Model (4). The variance function shows that the curves have much greater variation at the beginning and end of the interval of observation, and hence that the assumption of covariance stationarity about the mean function would be inappropriate. Figure 3 shows the order 13 correlation matrix for the sampled values plotted as a surface in which amount of correlation for two points in time is represented by the height of the surface. The smoothness of this surface is due to the fact that the functions giving rise to the sampled points are themselves highly regular. If the correlation r_{jk} between sampled values at points t_j and t_k is stationary, then it will be a function of $|t_j - t_k|$. This would imply that the surface in Figure 4 should fall off about the diagonal in the same manner for all points on the diagonal. In fact, this is approximately the case for the central third of the interval, but there is a conspicuous elevation and flatness of the surface at the beginning and ending points. In summary, both covariance and correlational stationarity are clearly violated for variation about a common mean curve $\mu(t)$.

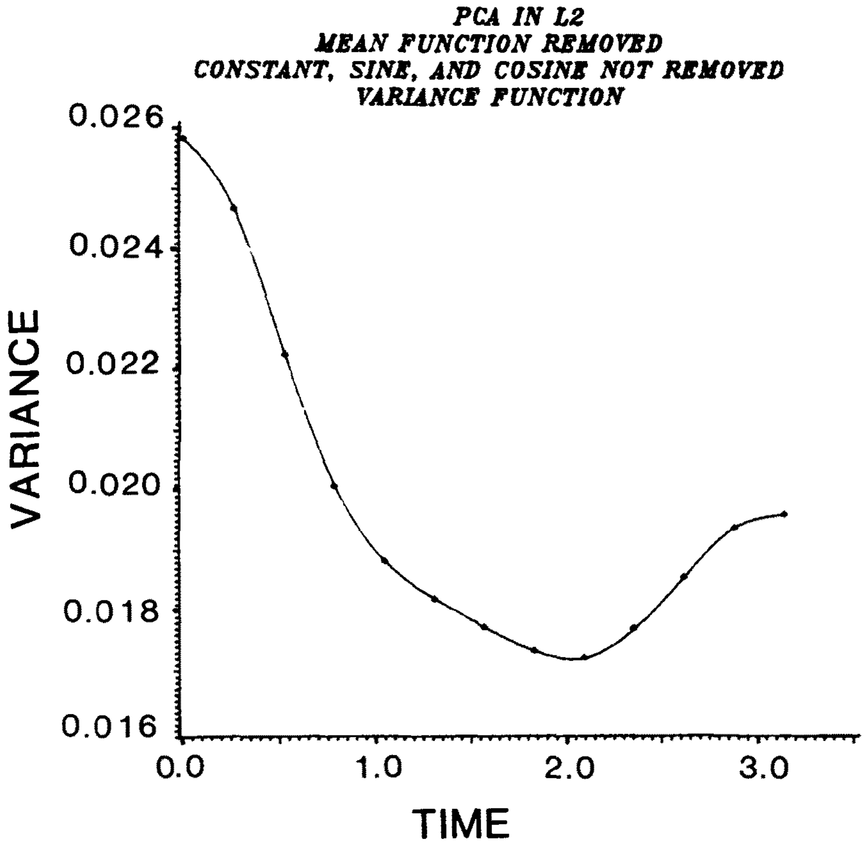

FIGURE 2
The variance of tongue dorsum position for each point in time. Points on the curve correspond to sampling points used in the data analyses reported in this paper.

The next step is to remove the trend due to (4) from each sampled curve. This was done by ordinary least squares regression on corresponding sampled values of the constant, sine, and cosine functions. The result is the matrix $X(I - P)$ where P is the order p projection matrix producing the least squares fit to any sampled curve in terms of these three functions. Figure 4 displays the variance functions for the residuals. It now has a strongly period character, but retains peaks at the initial and final points of the interval. Thus, the assumption of covariance stationarity of residuals produced by ordinary least squares analysis appears rather doubtful.

How revealing would classical principal components analysis (PCA) be for these data? Figure 5 shows the first three principal components of the covariance matrix for these data. In this paper we will follow the time series literature in referring to these components as harmonics. The first three harmonics account for all but 0.14% of the variation. The first harmonic accounts for record-to-record variation in overall tongue dorsum height and is well described by the constant function. The second describes a cosinusoidal component of variation even after the mean function in Figure 2 has been removed from the data. These two components are consistent with the ballistic motion hypothesis in (4). The third harmonic, accounting for only 1.6% of the variation, describes a further cosinusoidal component but with period π rather than 2π. Thus, PCA using ordinary unweight least squares tends to decompose variation into the first three compo-

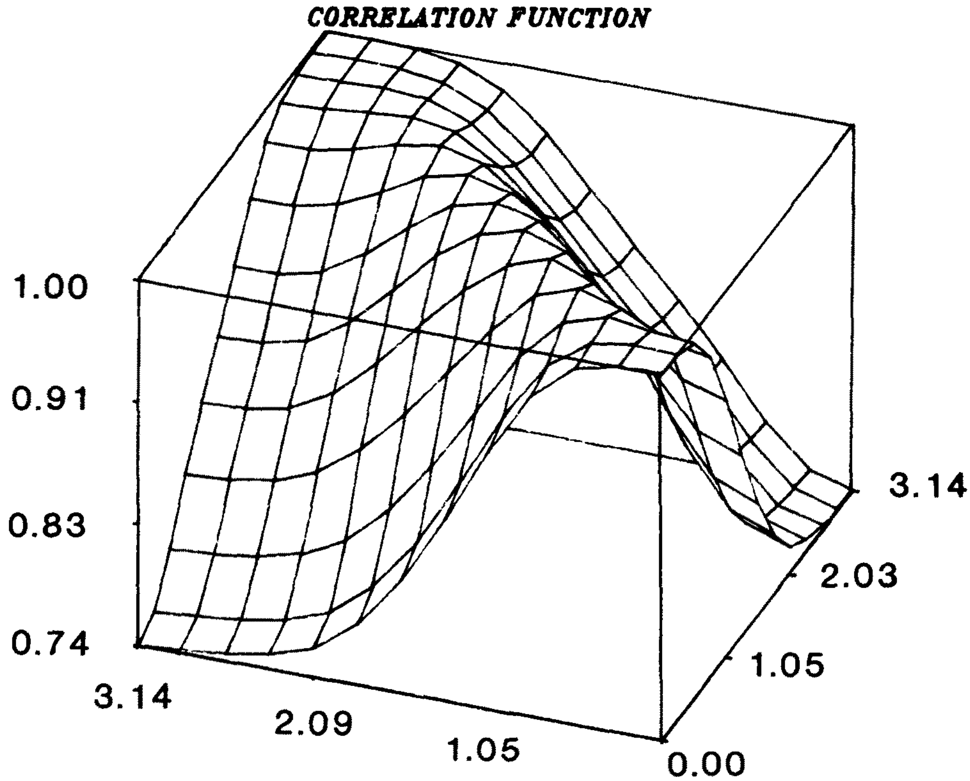

FIGURE 3
Correlations between tongue dorsum position for each pair of points in time. The height of the surface indicates the amount of correlation. Grid lines correspond to sampling points. Note the relative flatness of the surface at the initial and final points.

nents of an ordinary time series analysis. This is on the whole relatively unrevealing, and sheds little light on the ballistic motion hypothesis. Furthermore, it fails to highlight the clearly visible special effects at the interval endpoints.

A more powerful descriptive analysis of the data using conventional techniques is to remove the ballistic motion effects in (4) prior to PCA. This involves the eigenanalysis of the matrix $(I - P)X'X(I - P)$, from which the constant, sine, and cosine components have been removed by ordinary least squares. The first two harmonics now account for 97.9% of the residual variation, and these are displayed in Figure 6. These two harmonics both resemble cosine functions with periods π and $2\pi/3$, respectively. However, the first harmonic also seems to account for some variation at both endpoints, while the second does so for the final value. This gives some support to the hypothesis of ballistic motion in the central portion and external control at the endpoints, but the results at this point are less than striking. Both PCA analyses have begged the question of whether ordinary least squares analysis using the identity metric is really appropriate in this context.

4. Thinking About PCA in Function Space

The foregoing analyses, while certainly helpful, did not recognize in any way that the data arose from the observation of regular functions (the original polynomial splines). This, as we shall show below, is related to the use of ordinary least squares as opposed to generalized least squares with a weight matrix M. How can the regularity of the underly-

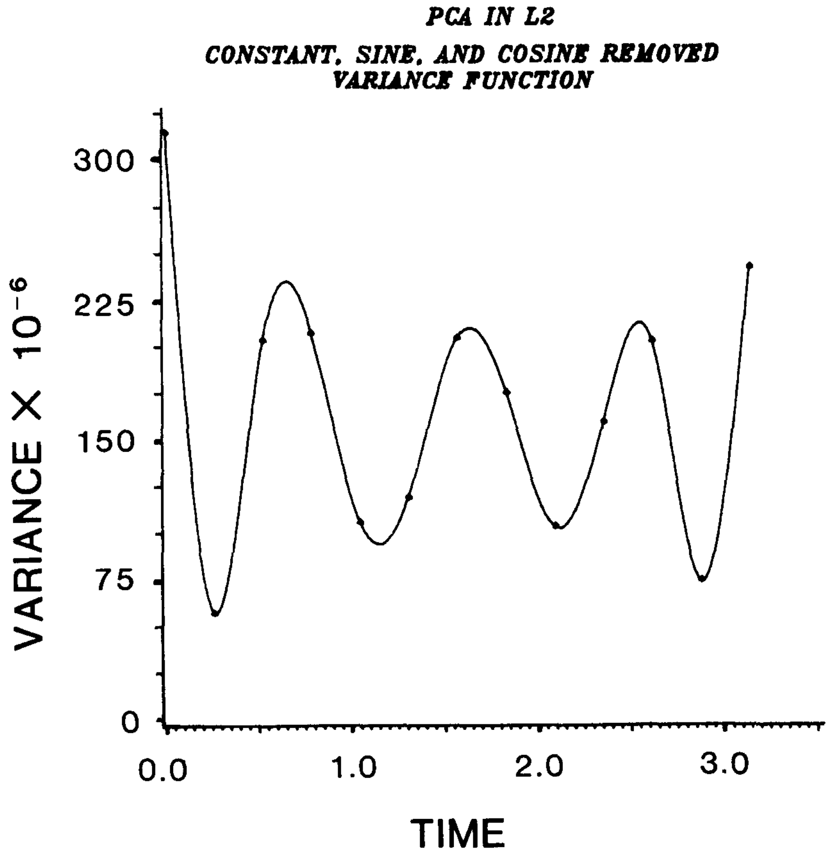

FIGURE 4
Variance of tongue dorsum position after the constant, sine, and cosine components of variation have been removed by ordinary least squares analysis.

ing functions be taken into account in the analysis of the sampled data? For example, is it possible to make use of information contained in the derivatives of the original functions as well as their values? Could PCA be expressed in a manner that would be independent of how many points are sampled or, indeed, whether the functions were sampled at all?

These objectives require a reformulation of the problem in terms of the functions themselves rather than their sampled values. In this section we pose the modeling and data analysis problems in terms of spaces of functions having a certain number of derivatives. We shall discuss the partitioning of this space into two components: one containing the model or known components of variation, and the other containing the residual variation. PCA analysis will then be expressed as a study of the variation of these functions either in the entire space or within either subspace.

We begin by assuming that that the sampled functions lie within the vector space $H^m(T)$ of functions defined on an interval $T = [a, b]$, possessing $m - 1$ absolutely continuous derivatives, and for which the square of the m-th derivative has a finite Lebesgue integral over T. This means that although the m-th derivative can be discontinuous at a countable number of points in T, lower order derivatives are not only continuous but differentiable. Any function x: $T \rightarrow R$ in this space can be represented as follows:

$$x = u + e, \tag{6}$$

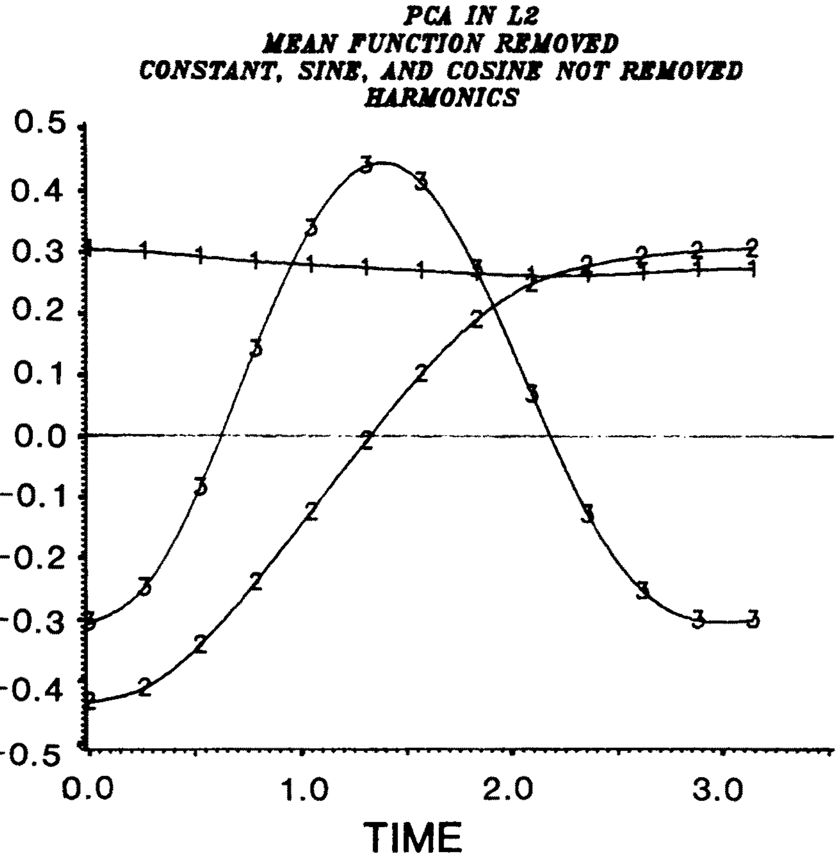

FIGURE 5
The first three components of variation resulting from a classical principal components analysis of the sample functions (means set to zero). These components account for 90.0%, 8.3%, and 1.5% of the variation, respectively, for a total of 99.8% of variance accounted for.

where u satisfies the homogeneous linear differential equation

$$Lu = \sum_{j=0}^{m} a_j D^j u = 0, \tag{7}$$

where the coefficients a_j may be continuous functions or constants, and $D^0 u = u$. It will be assumed for simplicity that a_m is nonzero. Decomposition (6) is motivated by the hypothesis that the functions will have a significant component within a certain class, which is determined by differential operator L. Functions within this class define an m-dimensional subspace of H^m. For example, if $L = D^m$, then $Lt^j = 0, j = 1, \ldots, m-1$, and this subspace is the space of polynomials of degree $m-1$. If $L = I + D$, then the subspace of dimension 1 is spanned by the function $\exp(-t)$. In this way, the appropriate choice of L can define the components u that carry the known or model component of the function space. We shall refer to this finite dimensional subspace of H^m as H_1. Alternatively, we may say that $H_1 = \ker(L)$.

The residual functions e are also within the space H^m. In order for the representation $x = u + e$ to be unique, it is essential to describe in some way that portion of H^m, which will be referred to as H_2, which does not contain elements in H_1. This can be done by imposing constraints on the possible functions e such that no element of H_1 could satisfy

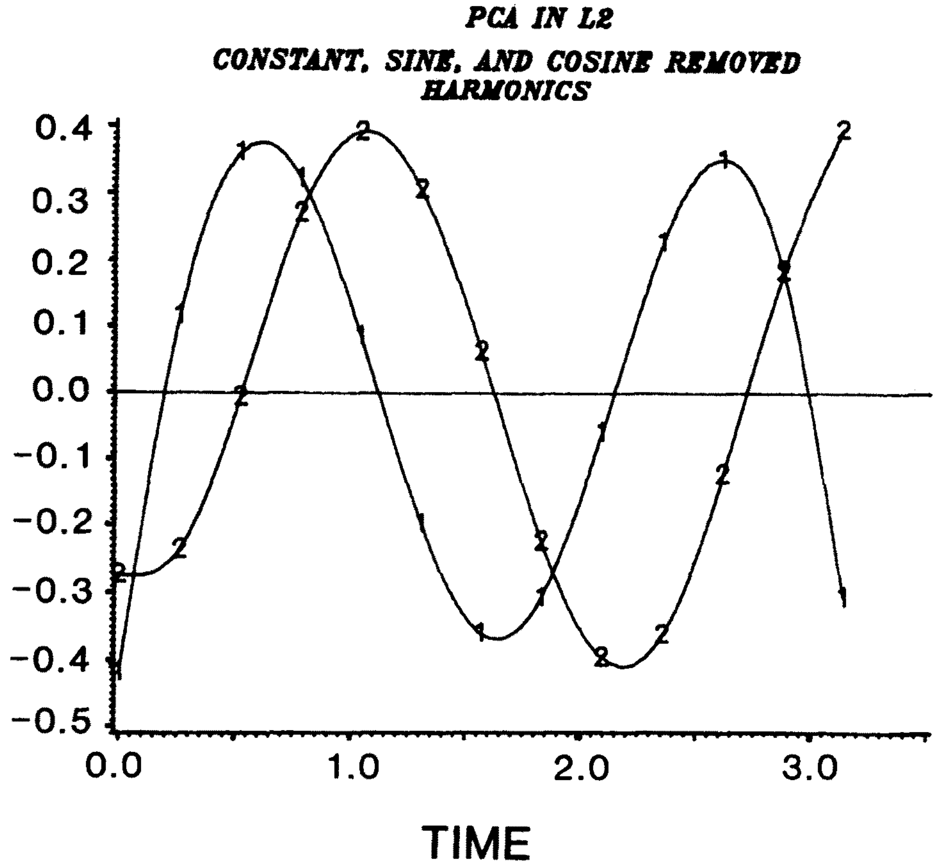

FIGURE 6

The first two components of variation from a classical principal components analysis of the data after removing the constant, sine, and cosine components. These components account for 70.3% and 27.6% of the residual variation, respectively, for a total of 97.6% of variance accounted for.

them. In this paper we shall impose *boundary constraints* on the values that elements of H_2 and their derivatives can take at certain points in the interval $T = [a, b]$. In general m such constraints will be required, and the particular choice employed will depend on what is convenient for the problem at hand. For example, for $L = I + D$ and the interval $T = [0, 1]$, the constraint $e(0) = 0$ will eliminate $\exp(-t)$ from H_2; while for $L = D^m$ and $T = [0, b]$ the constraints $D^j e(0) = 0, j = 0, \ldots, m - 1$ will eliminate all polynomials of degree $m - 1$ or less from H_2. The effect of these boundary constraints is to enable the expression of he function space as the direct sum

$$H^m = H_1 \oplus H_2. \tag{8}$$

In the case of the tongue data, for which the interval is $T = [0, \pi]$, we wish to model or possibly remove components which are linear combinations of 1, sin, and cos. Thus, the component u lies within the three-dimensional subspace H_1 of H^3 spanned by these components and satisfying the differential equation $Lu = Du + D^3u = 0$.

Procedures such as regression and principal components analysis presuppose an *inner product* defined on the vector space. In this case, H^m and its subspaces become Hilbert spaces when these inner products are defined, which we shall denote by $(\cdot, \cdot)_0$,

$(\cdot,\cdot)_1$, and $(\cdot,\cdot)_2$ for H^m, H_1, and H_2, respectively. A Hilbert space defined on the function space H^m is called a *Sobolev space.*

The inner product for H_2 which agrees naturally with its definition is

$$(x, y)_2 = \int LxLy\, dt. \tag{9}$$

Note that the boundary conditions imposed to define H_2 eliminate functions satisfying $Lx = 0$, so that $(x, x)_2 = 0$ if and only if $x = 0$. These same conditions suggest the following inner product for H_1:

$$(x, y)_1 = \sum_{j=1}^{m} B_j(x)B_j(y), \tag{10}$$

where the boundary functional $B_j(x)$ is the value determined by x for the j-th boundary condition. For example, if $L = I + D$ and we impose the boundary condition $e(0) = 0$, then $(x, y)_1 = x(0)y(0)$. Since the boundary conditions are chosen so as to exclude all elements of H_1 except for 0, we are assured that this inner product has the required positive definiteness on H_1. If we now define $(\cdot,\cdot)_0$ by

$$(x, y)_0 = (x, y)_1 + (x, y)_2, \tag{11}$$

then H_1 and H_2 are orthogonal subspaces, and there exist orthogonal projectors P_1 and P_2 onto H_1 and H_2, respectively, such that $(P_k x, P_k y)_0 = (x, y)_k$, $k = 1, 2$.

It now remains to define PCA in function space. Detailed discussions of the extension of PCA to arbitrary Hilbert spaces can be found in Besse (1979), Dauxois and Pousse (1976) and Dauxois, Pousse, and Romain (1982), and an elementary introduction is available in Ramsay (1982). Let us assume for simplicity that each of n functions x_i satisfies $\sum x_i(t) = 0$ for all $t \in T$. Then the extension of the notion of a covariance matrix to function space is the covariance function:

$$v(s, t) = n^{-1} \sum_{i=1}^{n} x_i(s)x_i(t), \qquad s, t \in T, \tag{12}$$

and corresponding to this function is the *covariance operator* defined by

$$\begin{aligned} Vx(s) &= \int v(s, t)x(t)\, dt \\ &= (v(s, \cdot), x). \end{aligned} \tag{13}$$

In conventional multivariate PCA one defines the principal axes to be the solution to the eigenequation $V\xi = \lambda\xi$, where $V = n^{-1}X'X$ and where any two eigenvectors are orthogonal. PCA can be defined more generally as the solution of this same eigenequation, where V in function space is understood to be defined by (13). The multivariate situation also has the generalization $VM\xi = \lambda\xi$ where any two eigenvectors ξ_j and ξ_k satisfy $\xi_j'M\xi_k = \delta_{jk}$. The positive definite matrix M determines the inner product in R^p; $(x, y) = x'My$. These two generalizations may be combined by using inner product notation to give the eigenproblem

$$(v(s, \cdot), \xi) = \lambda\xi(s), \qquad s \in T,$$

where in the multivariate case the index set T is $\{1, \ldots, p\}$.

Returning to the problem of PCA in function space, we see that there are three PCA analyses that may be interesting: PCA in H^m using $(\cdot,\cdot)_0$, PCA in H_1 which is in effect the PCA of the model components u, and PCA in H_2 which is the PCA of the residual components e. However, the objective of this paper is not to carry out PCA directly in any of these spaces, but to show how PCA in function space is equivalent to multivariate PCA in R^p with a particular choice of metric matrix M. In any case, it is not the PCA of arbitrary functions that interests us; instead we wish to analyze functions which *interpolate* the sampled points and are at the same time of minimum norm. In the next section we show how this is done.

5. Interpolation, Spline Functions, and Reproducing Kernels

With only sampled function values at our disposal, we seek some representative function h in H^m whose values at $t_1, \ldots, t_p$ will be "close" to these sampled values. However, even if by "close" one means to fit them exactly, there will be an infinite number of possible interpolants, and additional considerations are required. Since one would not wish an interpolant h to behave extravagantly between points being sampled, it is natural to require that it be as small as possible in either the sense $\|h\|_0$ or $\|h\|_2$. In the first case, the total behavior of h is required to be as close to the zero function as possible, while in the second the size of the residual e in the decomposition $h = u + e$ is required to be minimal thus ignoring the size of the model component u. A functional whose sampled values are equal to those observed and of minimal norm in either sense is called an *interpolating spline*. Alternatively, one may require that the composite loss function

$$Q(h) = \sum_{j=1}^{p} [x(t_j) - h(t_j)]^2 + \lambda \|h\|^2$$

be minimized for some parameter λ, with the norm being either for H^m or H_2. A minimizing function h with respect to such a loss is a *smoothing spline*. In order to reduce the amount of technical detail, we shall confine our attention in this paper to interpolating splines and will use only $\|h\|_0$ to define them. In many applications, including the classical examples of polynomial interpolating splines discussed in most textbooks, results will not depend on whether $\|h\|_0$ or $\|h\|_2$ is used.

The definition and computation of an interpolating spline when norms are defined as above using a linear differetial operator L requires one of two closely related concepts: the *Green's function* associated with L, or the *reproducing kernels* associated with the spaces H^m, H_1, and H_2. Since an exposition of the theory is both somewhat simpler and also somewhat more general within the context of reproducing kernels, we shall use this approach. The relationship between the Green's function and the reproducing kernel associated with L is discussed in the appendix.

5.1 *Basic Properties of Reproducing Kernels*

The reproducing kernel for a Hilbert space of functions defined on interval T is a bivariate function $k(\cdot, \cdot)$ defined on $T \times T$ which plays the same role as M^{-1} in a finite dimensional vector space with metric M. In a finite dimensional space $(m^i, x) = x_i$, where m^i is the i-th column of M^{-1}. Analogously, for a reproducing kernel Hilbert space of functions, k satisfies the basic *reproducing equation*

$$(k(s, \cdot), x) = x(s). \tag{15}$$

In this section the more useful properties of reproducing kernels are summarized. The reader is referred to treatments such as Aronszajn (1950), Aubin (1979), Duc-Jacquet (1973), and Shapiro (1971) for more details and proofs. Applications of reproducing kernels in other statistical contexts are to be found in Parzen (1961) and Kimeldorf and Wahba (1970, 1971).

In order to illustrate the concept of a reproducing kernel in a more familiar setting, let us begin by considering the space $L^2(T)$ of functions which are square-integrable. For any nonnegative kernel function $k(\cdot, \cdot)$ the integral transform $y(s) = \int k(s, t)x(t)\,dt$ has a smoothing effect, so that the transformation y is more regular and more spread out than the original function x. The amount of smoothing depends on how "spread out" the kernel function $k(\cdot, \cdot)$ is over R^2. In the extreme case of the Dirac delta functional, having the property $x(s) = \int \delta(s, t)x(t)\,dt$, the kernel has all of its mass concentrated on the diagonal values $\delta(t, t)$ and is zero elsewhere. As a consequence, it leaves the function x unchanged. Thus, it is the continuous analogue of the identity matrix. However, δ is not a function in the usual sense, and $\delta(s, \cdot)$ for fixed s is not a member of L^2. Rather, entities such as δ are more properly called generalized functions and the corresponding functional a distribution.

However, in spaces more regular than L^2 such as H^m there are symmetric functions $k(\cdot, \cdot)$ which behave like δ but which in addition are members of the same space when one argument is fixed.

Definition 1. For any open subset $T \in R$ and Hilbert space H of real functions defined on T, k: $T \times T \to R$ is called a *reproducing kernel* for E if: (a) $k(s, \cdot) \in H$ for all $s \in T$, (b) $(k(s, \cdot), x) = x(s)$ for all $s \in T$ and $x \in H$. If Hilbert space H possesses a reproducing kernel, it is called a *reproducing kernel Hilbert space.*

The following properties of reproducing kernels are demonstrated in the references cited above:

1. k is symmetric: $k(s, t) = k(t, s)$ for all $s, t \in T$.
2. k is positive in the sense that for any p-tuple $\{t_1, \ldots, t_p\}$ of elements in T the matrix K with elements $k(t_i, t_j)$ is positive semidefinite.
3. k is unique for a given space H.
4. The vector space generated by $\{k(s, \cdot), s \in T\}$ is dense in H.
5. If H is a direct sum $H_1 \oplus H_2$ of Hilbert spaces H_1 and H_2 with respective reproducing kernels k_1 and k_2 then $k(s, t) = k_1(s, t) + k_2(s, t)$.
6. The reproducing kernel for a finite dimensional Hilbert space H spanned by functions $u_1, \ldots, u_m$ is given by

$$k(s, t) = \sum_{i}^{m} \sum_{j}^{m} b^{ij} u_i(s) u_j(t), \tag{16}$$

where b^{ij} is the ij-th element of inverse of the matrix B containing values (u_i, u_j).

A Hilbert space of functions possesses a reproducing kernel only when the functions are sufficiently regular. More precisely, for any fixed value s the functional ρ_s: $H \to R$ which has the value $x(s)$ for argument x is continuous if and only if H is a reproducing kernel space. Property (b) above says that $\rho_s(x) = (k(s, \cdot), x)$. This ensures regularity since if the functions x_1 and x_2 are "near" in function space, then continuity of ρ_s implies that $x_1(s)$ and $x_2(s)$ are "near" to one another on the real line. The space L^2 does not satisfy this condition because of the possibility of discontinuities. Conversely, if one has a positive symmetric function k satisfying the conditions in Definition 1, then there exists a Hilbert function space H for which it is the reproducing kernel.

As Property 6 indicates, a finite dimensional Hilbert space always has a reproducing kernel. For example, if R^p has the inner product $(x, y) = x'My$, then the matrix M^{-1} understood as a real-valued mapping on $T \times T$, where T is the index set $\{1, \ldots, p\}$, has the above properties. In particular, $(m^i, x) = m^{i\prime}Mx = x_i$, where m^i is the i-th column of M^{-1}.

One way to study the characteristics of a reproducing kernel is to examine the properties of the functions $k_j = k(t_j, \cdot)$ defined by fixing one argument at each of the sampled values. These p functions play a central role in spline interpolation, as will be shown below. In the case of subspace H_1 spanned by the functions u solving $Lu = 0$, Property 6 implies that the k_j's will each be a particular linear combination of these functions u. Thus, for example, when $L = I + D^2$, the kernel of which is spanned by sin and cos, each k_j will be of the form $a_j \sin + b_j \cos$.

In the case of infinite dimensional space H_2, $k(t_j, \cdot)$ is still a piece-wise linear combination of a finite number of functions which depend on L. The inner product may be written

$$(x, y)_2 = (Lx, Ly)_{L^2}. \tag{17}$$

Associated with any linear operator is an *adjoint operator* L^* such that $(Lx, z) = (x, L^*z)$. The adjoint of $L = \sum a_j D^j$ is

$$L^*x = \sum (-1)^j D^j a_j x.$$

Since $(Lx, Ly) = (x, L^*Ly)$ and $(Lx, Lk_j) = (x, L^*Lk_j) = x(t_j)$, where the inner product is now in L^2, it follows that L^*Lk_j behaves like the Dirac delta function. Hence,

$$L^*Lk_j(t) = 0, \qquad t \neq t_j.$$

From this relation we can conclude that the reproducing kernel k_2 for H_2 will be a linear function of functions spanning $\ker(L^*L)$ in either argument, but will exhibit a discontinuity in terms of its derivative of order $2m - 1$ at the diagonal values $k_2(t, t)$. For example, consider the situation in which $L = D$ and $e(0) = 0$. Since $L^* = -D$, the reproducing kernel will have a zero second derivative everywhere except on the diagonal and thus be piecewise linear. Moreover, the first derivative of $k_2(s, t)$ in t will be discontinuous at $t = s$. In fact, it is simple to show that $k_2(s, t) = \min\{s, t\}$ and indeed has these properties.

Table 2 lists a number of Hilbert spaces along with their reproducing kernels which are of practical interest in the context of this paper. As a further specific example, consider the space of absolutely continuous continuous functions possessing a derivative in L^2 and such that $f(0) = 0$. Let the inner product for this subspace of H^1 be

$$(x, y)_2 = \int LxLy \, dt \qquad \text{where}$$

$$Lx = \lambda x + (1 - \lambda)Dx, \qquad 0 \leq \lambda < 1.$$

The closer λ is to unity, the more this inner product will approach that of L^2. The function satisfying $Lx = 0$ is $x = \exp[-\lambda t/(1 - \lambda)]$ and the boundary condition $f(0) = 0$ excludes this function from H_2, thus ensuring that (18) is an inner product. The reproducing kernel for H_2 is given by

$$k(s, t) = [\lambda(1 - \lambda)]^{-1} \exp(-\gamma s) \sinh(\gamma t), \qquad \gamma = \frac{\lambda}{(1 - \lambda)}, \qquad t \leq s, \tag{19}$$

and since $k(s, t)$ is symmetric, $k(s, t) = k(t, s)$ when $t > s$. Figure 7 displays the reproducing kernels for $\lambda = .5$ and $.9$. As λ approaches 1, note that k approaches δ in shape, which is consistent with the fact that the inner product approaches that for L^2. On the other hand, as λ approaches 0, the inner product approaches $\int DxDy \, dt$, and the reproducing kernel in this case is $k(s, t) = \min\{s, t\}$. Note that the function $k(s, \cdot)$ defined by fixing one of the arguments of k satisfies the boundary condition $k(s, \cdot) = 0$ and is an element of H_2. More-

Table 2. Some Examples of Inner Products and Associated Reproducing Kernels for $H^m[0,T]$

Space	Boundary Conditions	Inner Product	Reproducing Kernel $g(s,t)$ for $t \le s$
$\{1\}$		$u_0 v_0$	1
H^1	$u_0=0$	(Lu,Lv), $L=D$	t
$\{e^{-t}\}$		$u_0 v_0$	e^{-s-t}
H^1	$u_0=0$	(Lu,Lv), $L=I+D$	$e^{-s}\sinh(t)$
$\{1,t\}$		$u_0 v_0 + u_T v_T$	$1 - (s+t)/T + st/T^2$
$\{\sin,\cos\}$		$u_0 v_0 + u_T v_T$	$\{(1+C^2)\sin(s)\sin(t)$ $-SC[\sin(t)\cos(s)+\sin(s)\cos(t)] + S^2\cos(s)\cos(t)\}/S^2$
$\{\sin,\cos\}$		$u_0 v_0 + u_0' v_0'$	$\sin(s)\sin(t) + \cos(s)\cos(t)$
H^2	$u_0=u_T=0$	(Lu,Lv), $L=I+D^2$	$\{\sin(T-s)\sin(T-t)[2t-\sin(2t)]$ $+\sin(T-s)\sin(t)[\sin(s)\cos(T-s)$ $-\cos(s)\sin(T-s)+\cos(t)\sin(T-t)$ $-\sin(t)\cos(T-t)-2(s-t)C]$ $+ \sin(s)\sin(t)[2(T-s)-\sin(2T-2s)]\}/(4S^2)$
H^2	$u_0=u_0'=0$	(Lu,Lv), $L=I+D^2$	$[t\cos(s-t) - \cos(s)\sin(t)]/2$
$\{1,\sin,\cos\}$		$u_0 v_0 + u_0' v_0' + u_0'' v_0''$	$2-\cos(s)-\cos(t)+\sin(s)\sin(t)+\cos(s)\cos(t)$
H^3	$u_0=u_0'=u_0''=0$	(Lu,Lv), $L=D+D^3$	$t-\sin(s)-\sin(t)+\sin(s)\cos(t)-\cos(s)\sin(t)$ $+ \{[\sin(s)\cos(t)+\cos(s)\sin(t)][1-\cos(2t)]$ $+ \sin(s)\sin(t)[2t-\sin(2t)]$ $+ \cos(s)\cos(t)[2t+\sin(2t)]\}/4$

Notes: 1. $S = \sin(T)$, $C = \cos(T)$, $u_0 = u(0)$, $u_T = u(T)$, $g(t,s) = g(s,t)$

2. $T \neq k\pi$ for $L = I + D^2$, $D + D^3$, $u_0 = u_T = 0$

over, since $L^* = \lambda I - (1-\lambda)D$, $\ker(L^*L) = \text{span}\{\exp(\gamma t), \exp(-\gamma t)\}$, and $\sinh(\gamma t) = [\exp(\gamma t) - \exp(-\gamma t)]/2$, the result that $k(s,\cdot) \in \ker(L^*, L)$ is confirmed in this example.

The problem of how to calculate the reproducing kernel associated with a particular Hilbert Space is somewhat technical, and is taken up in the appendix.

5.2 Reproducing Kernels and Spline Interpolation

Reproducing kernels are valuable because they permit a simple account of the spline interpolation problem for reproducing kernel Hilbert spaces. Let H be such a space with reproducing kernel k and let $k_j = k(t_j, \cdot)$, $j = 1, \ldots, p$, be the p functions defined by fixing the first argument at the sampling points. Let us assume that the sampling points are such that these functions are linearly independent and thus span a subspace of H of dimension p. Let K be the symmetric matrix of order p containing the values of the reproducing kernel at the sampling points: $K := \{k(t_i, t_j); i, j = 1, \ldots, p\}$. Note also that $k_{ij} = (k(t_i, \cdot), k(t_j, \cdot))$ by the reproducing property of kernel k. For simplicity, it will be assumed that $t_1, \ldots, t_p$ have been chosen so that k is positive definite.

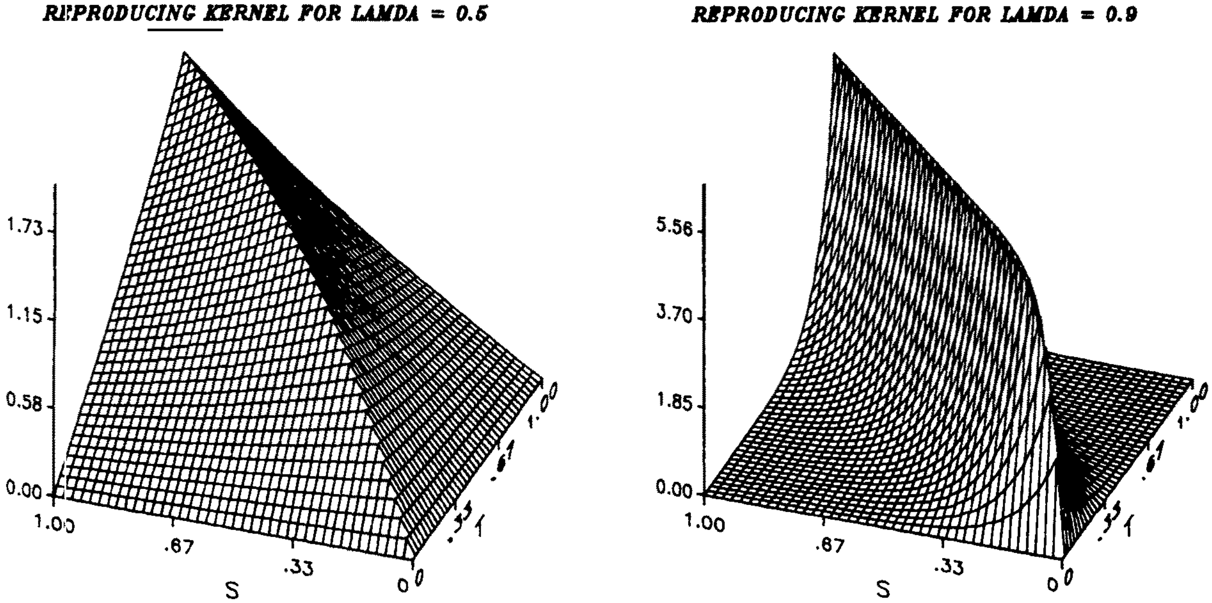

FIGURE 7
Reproducing kernels for the subspace H_2 associated with the differential operator $L = \lambda I + (1 - \lambda)D$ and the boundary condition $e(0) = 0$. The left function is for $\lambda = .5$ while the right is for $\lambda = .9$. Note that as λ approaches 1.0, the differential operators approaches the identity operator, and the reproducing kernel approaches the Dirac delta function δ.

Theorem 1. If K is positive definite then for any $x \in H$ there exists a unique interpolating function h such that $\|h\| \leq \|f\|$ for any interpolating function $f \in H$, and it is

$$h = \sum c_i k(t_i, \cdot) \quad \text{where} \quad c = K^{-1}x. \tag{20}$$

Proof. Since h interpolates x at the sampling points it satisfies the equations

$$(k(t_j, \cdot), h) = h(t_j) = x(t_j), \qquad j = 1, \ldots, p.$$

Substituting the above representation of h in these equations produces the linear system $\mathbf{x} = K\mathbf{c}$ where vectors $\mathbf{x}$ and $\mathbf{c}$ containing the interpolated values of $x(t)$ and the coefficients, respectively. The existence and uniqueness of h follow from the positive definiteness of K. Now let f be any interpolating function. The space H is the direct sum of the subspace span$\{k_j\}$ and its orthogonal complement. Since f interpolates $\mathbf{x}$, its projection on span$\{k_j\}$ is also h and thus it can be represented as

$$f = h + e, \qquad e \in \text{span}\{k_j\}, \quad \text{and} \quad \|f\|^2 = \|h\|^2 + \|e\|^2.$$

Since h does not depend on f, it follows that $\|f\|$ is minimized when $e = 0$. In matrix terms $X' = KC$ and the p by n matrix of the coefficients giving the minimum norm interpolants is $C = K^{-1}X'$. □

In general, the functions $k(t_i, \cdot)$ will not be orthogonal to one another, but they do provide a basis for the space of interpolating splines and are themselves splines. Thus, knowledge of the reproducing kernel k leads to a direct solution for the minimum norm interpolating spline h. For further material on splines Schumaker (1981) can be consulted and Wegman and Wright (1983) review other statistical applications of splines.

6. Principal Components Analysis of Interpolated Functions

6.1 PCA in H^m

We are now in a position to show the relationship between the PCA of the interpolated functions and the classical PCA of sampled values in a particular metric. This metric depends in a very simple way on the reproducing kernel for H, as the following theorem shows (Besse, 1979).

Theorem 2. When K is positive definite the PCA of interpolants $h_i \in H^m$ is equivalent to classical PCA of matrix X of sampled values in R^p in the metric K^{-1}.

Proof. The inner product of two interpolants h_1 and h_2 is

$$\begin{aligned}(h_1, h_2) &= (c_{j1}k_j, c_{j2}\,k_j)\\ &= c_1'Kc_2\\ &= \mathbf{x}_1'K^{-1}KK^{-1}\mathbf{x}_2\\ &= \mathbf{x}_1'K^{-1}\mathbf{x}_2,\end{aligned} \tag{21}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the vectors containing the sampled values for functions $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Thus, $(h_1, h_2) = (\mathbf{x}_1, \mathbf{x}_2)_R$ where the latter inner product is given by $\mathbf{x}_1'K^{-1}\mathbf{x}_2$. It follows that K^{-1} is the metric matrix for the classical PCA in R^p which is equivalent to the PCA of the interpolants in H^m. □

6.2. PCA in Subspaces H_1 and H_2 of H^m

This correspondence between PCA of interpolants h_i in H^m carries over to PCA in the subspace H_1 of model components u_i and the subspace H_2 of residuals e_i. In H_1 the goal is to explore the variation of the data in terms of functions chosen a priori and which satisfy the equation $Lu = 0$ for some differential operator L. Thus, analysis in H_1 is akin to a preliminary regression analysis on these components followed by a PCA of their fitted values. In H_2 the goal is to in effect remove from the analysis any variation in terms of these a priori model components in order to study any meaningful variation in the residuals.

The following theorem states the appropriate semimetric for the equivalent PCA in each case.

Theorem 3. The PCA of the interpolating spline functions in H_2 is equivalent to the classical PCA with semimetric $M_2 = K^{-1}K_2K^{-1}$.

Proof. The inner product of two interpolants h_1 and h_2 within H_2 is

$$\begin{aligned}(h_1, h_2)_2 &= c_1'K_2\,c_2\\ &= \mathbf{x}_1'K^{-1}K_2K^{-1}\mathbf{x}_2.\end{aligned} \tag{22}$$

Similarly, the semimetric matrix for the equivalence to PCA in H_1 is $M_1 = K^{-1}K_1K^{-1}$. In general M_1 and M_2 will not be of full rank.

Once a classical PCA in the appropriate metric has been completed, it is possible to return to function space since corresponding to the set of M-orthogonal eigenvectors ξ_j, $j = 1, \ldots, p$, satisfying

$$X'XM\xi_j = \lambda_j\,\xi_j$$

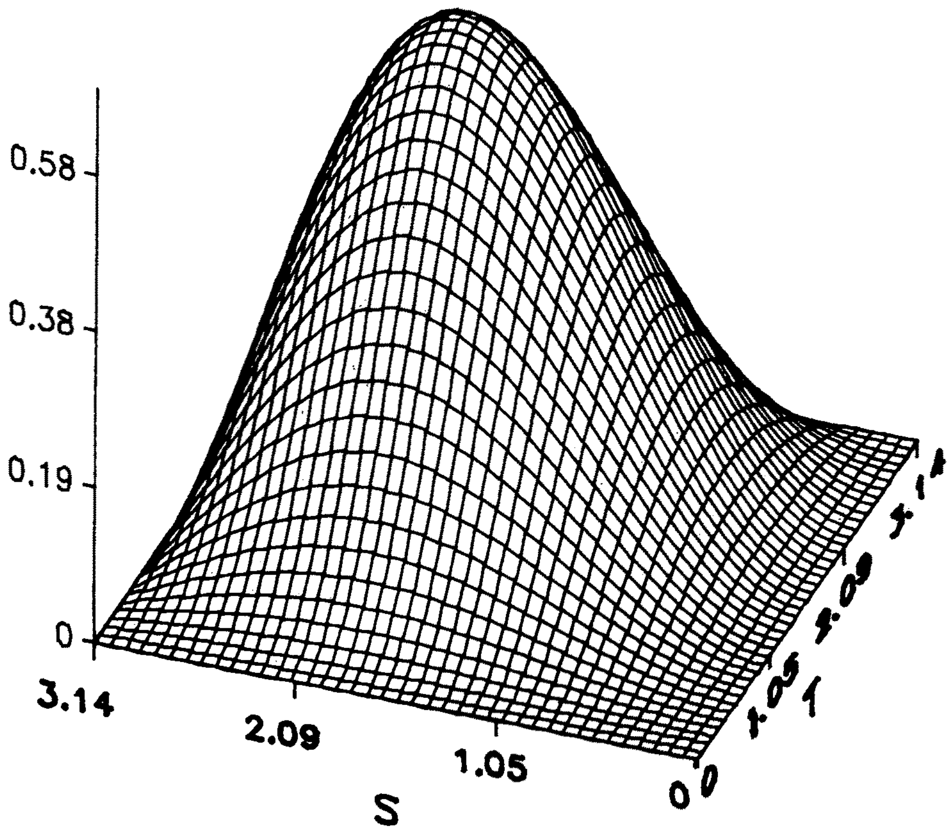

FIGURE 8
Reproducing kernel associated with the differential operator $L = D + D^3$ and the boundary conditions $e(0) = e'(0) = e(\pi) = 0$. This kernel is for the subspace H_2 of functions satisfying these constraints. Note that fixing either argument yields a function within this subspace, and hence satisfying these conditions.

is the set of spline interpolants of harmonics $e_j = \sum c_{jm} k_m$ where $c_j = K^{-1}\xi_j$. Thus, although the analysis can be carried out using the familiar machinery of matrix computation, the results can be expressed directly in functional analytic terms. □

7. A Functional Analysis of the Tongue Data

As we have seen, these sampled functions have strong components in the subspace spanned by the constant, sine, and cosine functions. The differential operator $L = D + D^3$ is thus appropriate in order to remove these components for purposes of studying the residual variation. In this section we present two analyses using this operator. These examples are designed to illustrate the importance of the boundary value constraints, $B_j(e) = 0, j = 1, 2, 3$. In the first example the choice of constraints is rather inappropriate to the nature of the variation, while in the second a better set of constraints leads to a more elegant description of the data. These examples also contrast two techniques for computing reproducing kernels.

7.1 Constraints $e(0) = De(0) = e(\pi) = 0$

This choice of boundary value constraints partitions H^3 into $H_1 \oplus H_2$, where H_1 consists of linear combinations of the constant, sine, and cosine functions, and H_2 consists

Table 3.
The First Five Eigenvalues for PCA's in H^3

	$e(0) = e(\pi) = e'(0) = 0$ Entire space	Constrained subspace	$e(\pi/4) = e(\pi/2) = e(3\pi/4) = 0$ Constrained subspace
1	.10	.18	1.98
2	.09	.09	.43
3	.05	.08	.02
4	.04	.04	.01
5	.04	.04	.001

of those functions whose initial and final values as well as initial slopes are zero. Thus, functions in H_2 may only vary in value within the interval, and may not vary in slope at zero. The main advantage of this choice of boundary constraints is that it is possible to work out an analytic expression for the Green's function associated with L and these constraints (see Appendix). This is

$$\begin{aligned} G(s; t) &= [1 + \cos(s)][1 - \cos(t)]/2 - \sin(s)\sin(t), \qquad t \leq s \\ &= -[1 - \cos(s)][1 + \cos(t)], \qquad s < t. \end{aligned} \tag{23}$$

The relation $k(s, t) = \int G(s; w)G(t; w)\,dw$ yields after some tedious but straightforward integration

$$\begin{aligned} k_2(s, t) = {} & \frac{(1 + C_s)(1 + C_t)(3t - 4S_t + S_t C_t)}{8} + \frac{S_s S_t(t - S_t C_t)}{2} \\ & - \frac{[S_t(1 + C_s) + S_s(1 + C_t)](2 - 2C_t - S_t^2)}{4} \\ & + \frac{S_s(1 - C_t)(2C_t - 2C_s + S_s - S_t^2)}{4} \\ & - \frac{(1 + C_s)(1 - C_t)(s - t - S_s C_s + S_t C_t)}{8} \\ & + \frac{(1 - C_s)(1 - C_t)(3\pi - 3s - 4S_s - S_s C_s)}{8}, \qquad t \leq s, \end{aligned} \tag{24}$$

where S_s, S_t, C_s, and C_t are sin (s), sin (t), cos (s), and cos (t), respectively. The value of $k_2(s, t)$ for $s < t$ can be obtained from the above expression by interchanging the roles of s and t because of the symmetry of the reproducing kernel. It can be verified that $k_2(s, t)$ satisfies the boundary conditions for all s, and that it is a piecewise combination of the functions 1, t, sin, cos, t sin, and t cos for any s. These functions are elements of ker $(L^*L) = \ker[-D^2(I + 2D + D^2)]$. Figure 8 displays the reproducing kernel k_2 as a surface, and its boundary value behavior is evident.

The reproducing kernel $k_1(s, t)$ corresponding to the inner product $(u, v)_1 = u(0)v(0) + u'(0)v'(0) + u(\pi)v(\pi)$ is given by (16) and is

$$k_1(s, t) = \frac{(1 + 2S_s S_t + C_s C_t)}{2}.$$

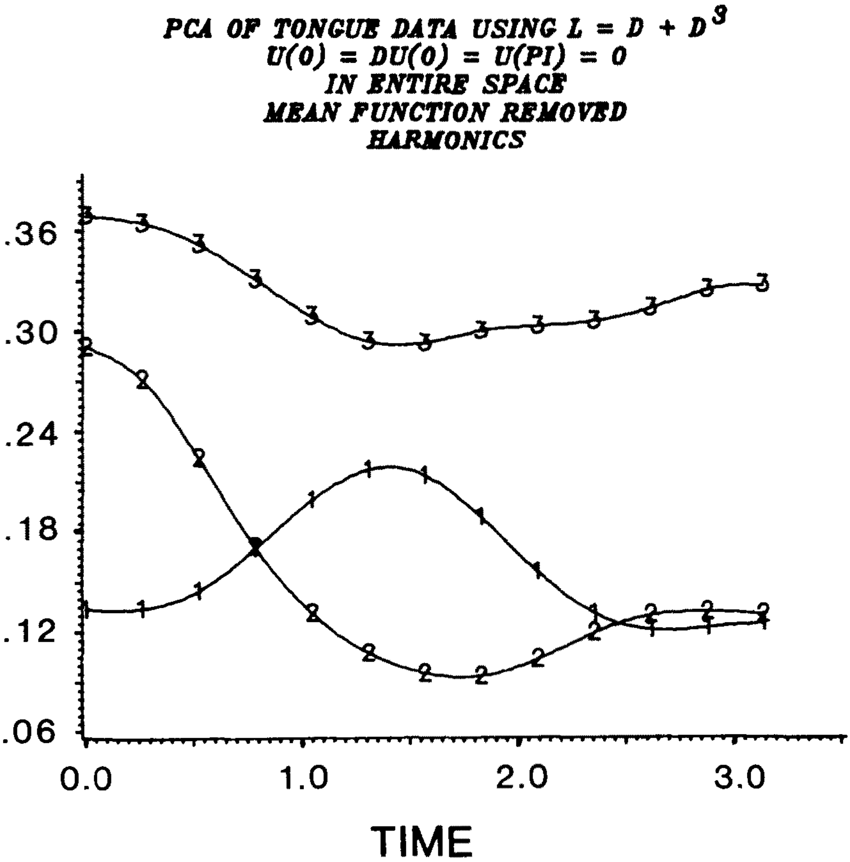

FIGURE 9
The first three components of variation resulting from a principal components analysis in H^3 with $L = D + D^3$ and boundary conditions $e(0) = e'(0) = e(\pi) = 0$. These components account for 25.8%, 22.7%, and 13.3% of the variation, respectively, for a total of 61.8% of the variance accounted for.

The first PCA with these boundary constraints is in the entire space H^3 using the inner product $(x, y)_0 = (x, y)_1 + \int LxLy\, dt$ with reproducing kernel $k_0 = k_1 + k_2$. For this analysis the metric is $M = (K_1 + K_2)^{-1}$, where matrices K_1 and K_2 are formed from the respective reproducing kernels evaluated at each of the pairs of sampling points. Table 3 gives the first five eigenvalues of the eigenequation $X'XM\xi = \lambda\xi$, where data matrix X has column means of zero. These eigenvalues descend very gradually indicating that a rather large number of eigenfunctions or harmonics would be required to approximate the data well in this metric. The first three eigenfunctions or harmonics are displayed in Figure 9. These give a rather different image of the data than those generated from the L_2 analysis in section 3. In particular, the harmonics emerge in the opposite order, with the dominant harmonic displaying variation in the central region of the interval and having the appearance of a sinusoid with period π. The reason for this is that the inner product $(\cdot, \cdot)_0$ is much more sensitive to the contribution of $(\cdot, \cdot)_2$ than that for $(\cdot, \cdot)_1$, while the L_2 analysis of necessity pays no attention to derivative information.

The second PCA analyzes variation which is only in H_2 through the use of the metric $M = (K_1 + K_2)^{-1}K_2(K_1 + K_2)^{-1}$. In effect, this analysis combines the regression analysis phase and the PCA of residuals which was employed in the second least squares analysis.

The first five eigenvalues are also given in Table 3, and do not differ greatly from these in the first analysis. The first two eigenfunctions must satisfy the boundary constraints, and thus display only variation in the central portion of the interval.

These results are disappointing in two respects. First, the gradual descent of the eigenvalues implies that a large number of eigenfunctions would be required, and thus an accurate description of the data would hardly be very parsimonious. Secondly, the dominant eigenfunctions do not have any obvious interpretation in terms of the tongue dorsum movement. The cause of these problems is clear from Figure 3, where it can be noted that the system deviates from ballistic motion primarily at the endpoints of the interval. That is, stripping off the constant, sine, and cosine components which would characterize pure harmonic motion should show strong residuals near the endpoints. Unfortunately, the boundary constraints employed here do not permit any such variation in H_2 in these regions. In effect, we are fitting the functions u in H_1 exclusively to the endpoints and the initial slope, which is precisely where we expect these components are least relevant. Thus, the conventional L_2 analysis was more useful in that these components were removed by least squares, which is less sensitive to endpoint variation.

7.1 *Constraints* $e(\pi/4) = e(\pi/2) = e(3\pi/4) = 0$

These constraints partition H^3 into $H_1 \oplus H_2$, where H_1 now has the inner product

$$(u, v)_1 = u\left(\frac{\pi}{4}\right)v\left(\frac{\pi}{4}\right) + u\left(\frac{\pi}{2}\right)v\left(\frac{\pi}{2}\right) + u\left(\frac{3\pi}{4}\right)v\left(\frac{3\pi}{4}\right),$$

and H_2 now consists of functions which are zero at these three points. These constraints both partition H^3 into the desired components and permit elements of H_2 to vary at the endpoints as we wish. In effect, the components $u \in H_1$ are determined so as to fit the functions at $\pi/4$, $\pi/2$, and $3\pi/4$, which is within the region where the tongue movement appears to exhibit simple harmonic motion.

We move directly to the analysis of the data in H_2. The reproducing kernel for H_2 has defied our attempts to express it analytically, but fortunately reasonable approximation procedures are available for computing the values of $k_2(t_i, t_j)$. These are briefly described in the appendix. The dominant eigenvalues for this analysis are displayed in Table 3. Now we see that there are only two large eigenvalues, which account for 98.5% of the variation. The two corresponding eigenfunctions are displayed in Figure 10. Their interpretation is obvious: the first accounts for a departure from simple harmonic motion which is a simultaneous deviation in the same direction at the two endpoints, while the second harmonic describes the extent to which the tongue is too low initially and too high finally (or vice versa). Thus we arrive at a result which displays clearly the effects of neural input at the highest and lowest points of the tongue's trajectory. Moreover, the fact that all remaining eigenvalues are very small suggests that simple harmonic motion describes tongue dorsum behavior very adequately in the intermediate region of $[0, \pi]$.

Although these results are clearly more useful than those using the endpoint boundary constraints, how much better are they than the L^2 results? The results differ primarily in terms of interpretation. In L^2 the leading two dimensions are in H_1, and the image in H_2 would lead to supposing that residual motion was also sinusoidal but of a higher frequency. The second H^3 analysis, however, explicitly takes into account the possibility that residual variation is primarily at the endpoints, which is clearly evident in the variance and covariance plots. Although residual sinusoidal motion of period π was not excluded from this analysis, nevertheless the results confirmed that the tongue departs from ballistic motion primarily at the endpoints, and probably in response to bursts of input. Experiments are now being planned to confirm this by direct observation.

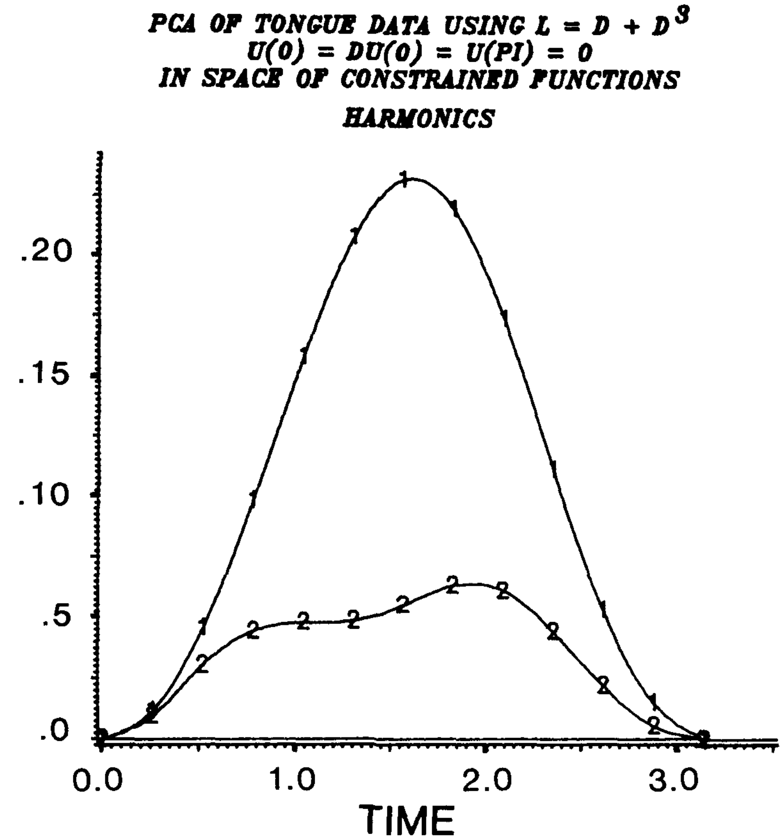

FIGURE 10
The first two components of variation resulting from the analysis used in Figure 9 but in the space H_2 of constrained functions. These account for 36.3% and 17.8% of the variation, respectively, for a total of 54.1% of the variance accounted for.

8. Conclusions and Summary

Data collected in the behavioral sciences often involve time as an independent variable. To be sure, such data do not usually arise from observing a process at every instant in time, and instead consist in observations at a limited number of discrete sampling points. This does not mean, however, that the study of such data cannot be adapted to take into account the behavior of first and higher order derivatives in the underlying process. We have shown here that taking derivative information into account amounts to a change of metric for classical multivariate techniques.

The appropriate metric for the analysis of temporal (or other sampled function) data has a very natural expression in terms of the reproducing kernel. The reproducing kernel in turn is intimately related to the concept of a spline function. Although the mathematical technology associated with these concepts may be somewhat unfamiliar to most data analysts trained in the classical tradition, we hope that we have provided an interesting argument for becoming familiar with these notions.

Finally, a number of generalizations of the results in this paper will be left to subsequent publications. The problems of vector-valued functions of time and of real- and vector-valued functions of vector arguments involve comparatively minor modifications of the structures described here. The problem of noise or random variation in the data has

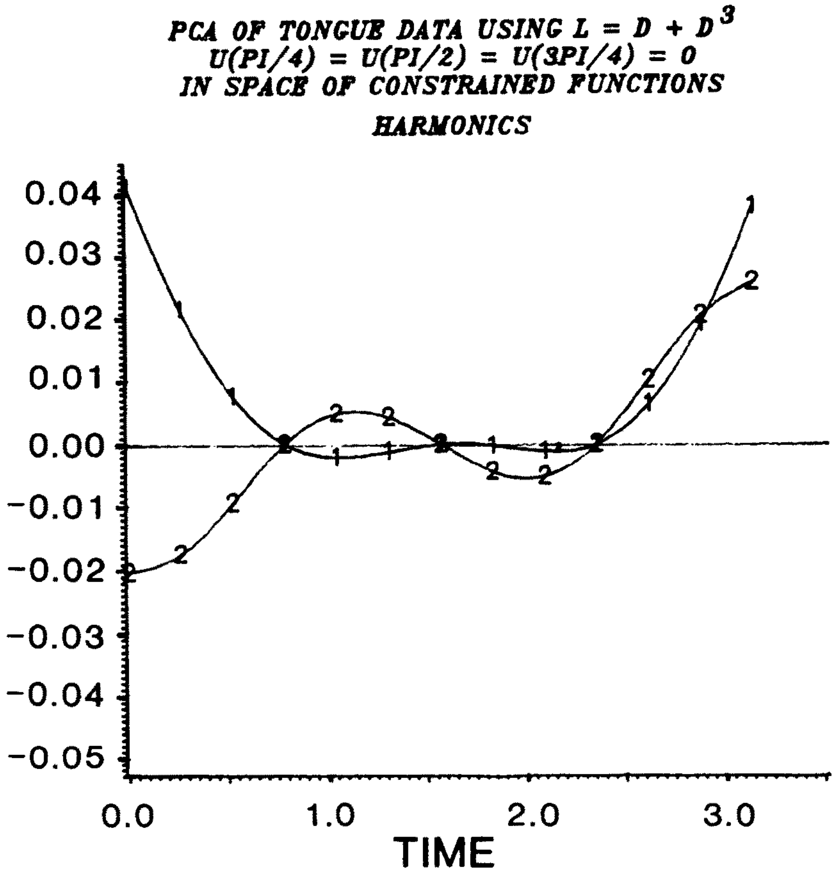

FIGURE 11
The first two components of variation resulting from an analysis using $L = D + D^3$ and boundary conditions $e(\pi/4) = e(\pi/2) = e(3\pi/4) = 0$. The analysis is in the space H_2 of constrained functions. The components account for 80.8% and 17.7% of the variation, respectively, for a total of 98.5% of the variance accounted for.

also not been dealt with, but is of obvious importance. We expect that there is much to learn concerning the appropriate choice of inner product and computation of the associated reproducing kernel for particular applications.

Appendix: Computing Reproducing Kernels

There are two approaches described here to computing the reproducing kernel k_2 associated with H_2, where H_2 is a space of functions satisfying certain constraints and having the inner product $\int LxLy\, dt$. The first approach involves calculating the reproducing kernel analytically using the Green's function $G(s; t)$ associated with the operator L and the boundary conditions. Once the Green's function is available, it is a routine matter to compute the reproducing kernel. The second approach involves approximating the matrix $K_2 = k_2(t_i, t_j)$ by using B-spline approximations to the Green's function and to the functions in H_2. This approach is attractive when the constraints are such as to make the computation of the analytic solution intractable. In both approaches it is assumed that the number of constraints equals the degree of the linear operator and that they exclude functions in $\ker(L)$.

A1. Analytic Solution for the Reproducing Kernel

The Green's Function $G(s;\, t)$ associated with L and a set of boundary value constraints is a bivariate function which provides the inverse of the operation Lu in the following sense:

$$u(s) = \int G(s;\, t) Lu(t)\, dt, \tag{A1}$$

where u is assumed to satisfy the constraints. The properties of Green's functions are described in detail in Roach (1982) and Stakgold (1979). These properties may be summarized as follows:

1. $LG(\cdot\,;\, t) = 0$ for any fixed t.
2. $G(\cdot\,;\, t)$ satisfies the boundary value contraints for any fixed t.
3. If $m > 1$, then $D^j G(s;\, s^+) = D^j G(s;\, s^-)$ for any fixed $s, j = 0, \ldots, m - 2$.
4. $D^{m-1}G(s;\, s^+) - D^{m-1}G(s;\, s^-) = 1$ for any fixed s.

Condition 1 implies that as a function of s G lies in the kernel of L and hence that G is of the form

$$\begin{aligned} G(s;\, t) &= f_1(s)a_1(t) + \cdots + f_m(s)a_m(t), \qquad t \leq s, \\ &= f_1(s)b_1(t) + \cdots + f_m(s)b_m(t), \qquad s \leq t, \end{aligned} \tag{A2}$$

where $f_1, \ldots, f_m$ are known functions spanning the kernel of L and $a_1, \ldots, a_m, b_1, \ldots, b_m$ are unknown. Thus, there are $2m$ unknown functions to be determined.

The simplest procedure for determining G is to use the m boundary conditions in Property 2, the $m - 1$ continuity conditions in Property 3, and the discontinuity condition in Property 4 to determine these unknown functions. The following example illustrates how this process works. Let $L = D + D^3$ for the space $H^3[0, \pi]$ with the boundary conditions $u(0) = u'(0) = u(\pi) = 0$. The kernel of L is spanned by the functions $\{1, \sin, \cos\}$ and thus the Green's function is of the form

$$\begin{aligned} G(s;\, t) &= a_1(t) + a_2(t) \sin(s) + a_3(t) \cos(s), \qquad 0 \leq t \leq s, \\ &= b_1(t) + b_2(t) \sin(s) + b_3(t) \cos(s), \qquad s \leq t \leq \pi. \end{aligned} \tag{A3}$$

However, the boundary condition $G(0;\, t) = 0$ implies $b_1(t) + b_3(t) = 0$, $G'(0;\, t)$ implies $b_2(t) = 0$, and $G(\pi;\, t) = 0$ implies $a_1(t) - a_3(t) = 0$. Thus, we can simplify $G(s;\, t)$ to the following form:

$$\begin{aligned} G(s;\, t) &= a(t)(1 + \cos(s)) + c(t) \sin(s), \qquad t \leq s, \\ &= b(t)(1 - \cos(s)), \qquad s \leq t. \end{aligned} \tag{A4}$$

The continuity of $G(s;\, \cdot)$ and $DG(s;\, \cdot)$ imply the two equations

$$\begin{aligned} &a(s)(1 + \cos(s)) - b(s)(1 - \cos(s)) + c(s) \sin(s) = 0 \\ &a'(s)(1 + \cos(s)) - b'(s)(1 - \cos(s)) + c'(s) \sin(s) = 0 \end{aligned} \tag{A5}$$

while the discontinuity condition implies

$$a''(s)(1 + \cos(s)) - b''(s)(1 - \cos(s)) + c''(s) \sin(s) = 1. \tag{A6}$$

A little manipulation of these linear differential equations yields the solutions

$$a(s) = \frac{(1 - \cos(s))}{2}, \qquad b(s) = \frac{-(1 + \cos(s))}{2}, \qquad c(s) = -\sin(s). \tag{A7}$$

The relation between the reproducing kernel and the Green's function is as follows:

$$G(s; \cdot) = Lk(s, \cdot) \qquad \text{for any fixed } s. \tag{A8}$$

To see this, note that $(Lk(s, \cdot), Lu) = (G(s; \cdot), Lu) = u(s)$ and thus that $Lk(s, \cdot)$ is the kernel of the integral transform reversing the effect of the differential operator L. That it satisfies the remaining properties of the Green's function is obvious. Moreover,

$$(G(s; \cdot), G(t; \cdot)) = (Lk(s, \cdot), Lk(t, \cdot)) = k(s, t), \tag{A9}$$

so that it suffices to know the Green's function associated with the differential operator L and the boundary conditions to determine the reproducing kernel.

An alternate route to determining the Green's function analytically which is more general but more tedious is to note that

$$(Lk(s\cdot,), Lu) = (L^*Lk(s\cdot,), u) = u(s), \tag{A10}$$

where L^* is the adjoint of the operator L. This implies that

$$L^*Lk(s, \cdot) = L^*G(s; \cdot) = \delta(s, t) \tag{A11}$$

from which we can deduce that $G(s; \cdot)$ lies in the kernel of L^* for any fixed s and $t \neq s$. Thus, in the equations (4) to (7) we can represent the functions $a(t)$, and $c(t)$ as $a_1 + a_2 \sin(t) + a_3 \cos(t)$, and etc. since $L^* = -L$ for $L = D + D^3$, and L^* has the same kernel. This leaves the nine constants $a_1, \ldots, c_3$ to be determined so as to satisfy the continuity and discontinuity conditions.

A2. Approximating the Matrix of Reproducing Kernel Values

In the previous section it was noted that a value of the reproducing kernel is given by

$$k_2(t_i, t_j) = \int G(t_i; t)G(t_j; t)\, dt.$$

This suggests that if one could approximate the function $G(t_i; \cdot)$ as a linear combination of known functions for each value of t_i, then the rest would be a matter of numerical or analytic integration of the approximation. The family of B-spline functions are admirably well suited to this purpose since a B-spline can be converted to a polynomial at any argument value, and hence the product of two B-splines can be easily integrated. Moreover, B-splines are a very powerful set of approximating functions. Our concern here, of course, is not with the quality of the approximation to $G(s; t)$, but rather the quality of the resulting estimate of $k_2(t_i, t_j)$. The integration required to get this value will tend to smooth out any local errors in the estimates of the two Green's functions.

B-splines are also well-suited to approximating elements in H_2 since it is a relatively simple matter to constrain their values or the values of their derivatives at specified points through the appropriate choice of knots.

In order to approximate the univariate function $G(t_i; \cdot)$, let a knot sequence t_{1q}, $q = 1, \ldots, N_1 + I_1$ be chosen, where I_1 is the order of the B-splines to be used. In general this sequence will have the first I_1 knots equal to 0, and the final I_1 knots equal to T, where the $t \in [0, T]$. The remaining knots can be equally spaced provided that N_1 is sufficiently large. Further restrictions on N_1 will be given below. This sequence determines a set of N_1 B-splines $B_{1q}(t)$, and the approximation to the Green's function is

$$G(t_i; t) = \sum_q a_{iq} B_{1q}(t).$$

The problem then becomes how to choose the coefficients a_{iq} properly. If we let $G(t)$ be the p-dimensional function $\{G(t_i; t)\}$ and $B_1(t)$ be the N_1-dimensional function $\{B_{1q}(t)\}$, then we have that

$$G(t) = AB_1(t),$$

where the coefficient matrix A is p by N_1.

In order to approximate H_2 by the span of a finite number N_2 of B-splines, let a knot sequence t_{2q}, $q = 1, \ldots, N_2 + I_2$ be chosen, where again I_2 is the order of the B-splines to be used. This knot sequence must be such as to ensure that each of the associated B-splines $B_{2q}(t)$, $q = 1, \ldots, N_2$ satisfy the required constraints. In particular, if one requires that $B_{2q}(c) = 0$ for some $c \in [0, T]$, then I_2 of the knots must equal C, and this will result in a single B-spline being nonzero at this point. This spline is then eliminated from the final sequence of B-splines. Constraints on derivatives can also be handled by this technique. Thus, since some of the B-splines associated with the knot sequence are to be dropped, we will have in general that $N_2 < N_2'$. Again let $B_2(t)$ be the N_2-dimensional function $\{B_{2q}(t)\}$. By using the relation

$$\begin{aligned} B_{2q}(t_i) &= \int G(t_i, t) LB_{2q}(t)\, dt \\ &\approx \int \sum_r a_{ir} B_{1r}(t) LB_{2q}(t)\, dt, \qquad i = 1, \ldots, p, \end{aligned}$$

we arrive at the matrix equation $Y \approx AX$ where the p by N_2 matrix Y is given by

$$y_{iq} = B_{2q}(t_i),$$

and the N_1 by N_2 matrix X is given by

$$x_{rq} = \int B_{1r}(t) LB_{2q}(t)\, dt.$$

Since $LB_{2q}(t)$ is also a B-spline, but of order $I_2 - m$, we must have that $I_2 > m$. The integration required to compute x_{rq} can be carried out analytically by converting the two B-splines to polynomials within each inter-knot interval.

The coefficient matrix A can then be computed by standard least squares procedures provided that rank $(X) \leq N_2$. This implies in general the restriction $N_1 \geq N_2$.

Once the coefficient matrix A is obtained, the reproducing kernel matrix K_2 is then approximated by AZA', where matrix Z of order N_1 has elements

$$z_{qr} = \int B_{1q}(t) B_{1r}(t)\, dt,$$

where these are computed by the same techniques used to compute the elements of X.

Experience to date indicates that choosing N_1 and N_2 to be of the order of 100 gives very good approximations to K_2. In practice N_1 and N_2 can be increased until no appreciable change in any element of K_2 occurs. On the other hand, there appears little to be gained by using very high order B-splines.

References

Anderson, T. W. (1971). *The statistical analysis of time series.* New York: Wiley.

Aronszajn, N., (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68,* 337–404.

Aubin, J.-P. (1979). *Applied functional analysis.* New York: Wiley Interscience.

Besse, P. (1979). Etude descriptive des processus: Approximation et interpolation [Descriptive study of processes: Approximation and Interpolation]. Thèse de 3 cycle, Université Paul-Sabatier, Toulouse, 1979.

Dauxois, J., & Pousse, A. (1976). Les analyses factorielles en calcul des probabilité et en statistique: Essai d'étude synthètique [Factor analysis in the calculus of probability and in statistics]. Unpublished doctoral dissertation, l'Université Paul-Sabatier de Toulouse, France.

Dauxois, J., Pousse, A., & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis, 12*, 136–154.

Duc-Jacquet, M. (1973). Approximation des fonctionelles lineaires sur des éspaces hilbertiens auto-reproduisants [Approximation of linear functionals on reproducing kernel Hilbert spaces]. Unpublished doctoral dissertation, Grenoble.

Hunter, I. W., & Kearney, R. E. (1982). Dynamics of human ankle stiffness: Variation with mean ankle torque. *Journal of Biomechanics, 15*, 747–752.

Huxley, A. F. (1980). *Reflections on Muscle.* Princeton: Princeton University Press.

Keller, E., & Ostry, D. J. (1983). Computerized measurement of tongue dorsum movements with pulsed-echo ultrasound. *Journal of the Acoustical Society of America, 73*, 1309–1315.

Kimeldorf, G. S., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics, 41*, 495–502.

Kimeldorf, G. S., & Wahba, G. A. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Application, 33*, 82–94.

Munhall, K. G. (1974). Temporal adjustment in speech motor control: Evidence from laryngeal kinematics. Unpublished doctoral dissertation, McGill University.

Parzen, E. (1961). An approach to time series analysis. *Annals of Mathematical Statistics, 32*, 951–989.

Ramsay, J. O. (1982). When the data are functions. *Psychometrika, 47*, 379–396.

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

Rao, C. R. (1964). The use and interpretation of principal components analysis in applied research. *Sankhya, 26*(A), 329–358.

Rao, C. R. (1980). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In P. R. Krishnaiah (Ed.), *Multivariate analysis V* (pp. 3-22). Amsterdam: North-Holland.

Roach, G. F. (1982). *Green's Functions*, Cambridge: Cambridge University Press.

Schumaker, L. (1981). *Spline Functions: Basic Theory*, New York: Wiley.

Shapiro, H. S. (1971). *Topics in Approximation Theory.* New York: Springer-Verlag.

Stakgold, I. (1979). *Green's Functions and Boundary Value Problems.* New York: Wiley.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23*, 19–23.

Wegmen, E. J., & Wright, I. W. (1983). Splines in statistics. *Journal of the American Statistical Association, 78*, 351–365.

Manuscript received 2/28/84.
Final version received 12/10/85.