PSYCHOMETRIKA—VOL. 47, NO. 4.
DECEMBER, 1982

WHEN THE DATA ARE FUNCTIONS

J. O. RAMSAY

MCGILL UNIVERSITY

A datum is often a continuous function $x(t)$ of a variable such as time observed over some interval. One or more such functions are observed for each subject or unit of observation. The extension of classical data analytic techniques designed for p-variate observations to such data is discussed. The essential step is the expression of the classical problem in the language of functional analysis, after which the extension to functions is a straightforward matter. A schematic device called the duality diagram is a very useful tool for describing an analysis and for suggesting new possibilities. Least squares approximation, descriptive statistics, principal components analysis, and canonical correlation analysis are discussed within this broader framework.

Key words: continuous data, functional analysis, duality diagram.

Introduction

Sophisticated data collection hardware often produce data which are a set of continuous functions. I am sure that all of us have seen such data: EEG and EMG records, learning curves, paths in space, subject responses continuous in time, speech production measurements during vocalization, bioassay data, and so on. Consider as a further example the curves displayed in Figure 1. These indicate the height of the tongue dorsum during ten different utterances of the sound "ah-kah" by a single subject [Keller & Ostry, Note 1]. It is natural to consider each curve as a single observation, to summarize the ten curves in terms of an average curve, and to measure in some way the variation of the curves about this average.

This paper considers the extension of classical statistical techniques to include functional data. It will be an elementary and simplified treatment, which may annoy those wanting more subtlety and rigor. I must warn you, however, that a fundamental change of point of view about what data are will be required, and if you leave my address aware that an altered state of statistical consciousness is possible, I shall be content.

In dealing with functional data I will refer frequently to two lines of development. The first is the expression of traditional data analytic technology in the language of functional analysis. Much of this work has taken place in France and is not available in English. I am particularly indebted to the mongraphs of Cailliez and Pagès [1976] and Dauxois and Pousse [1976]. We are very fortunate to have with us for these meetings a number of those associated with this work, and in part my talk is only an introduction to tomorrow's symposium.* The second line of development that has fascinated my colleague Suzanne Winsberg and I in recent years has been statistical applications of spline functions. I feel that splines are destined to play a fundamental role in the analysis of functional data, but I will try to show how in only a vague way at this point. Finally, this

* "New glances at principal components and correspondence analysis" was a symposium at the 1982 Joint Meetings of the Classification Society and Psychometric Society, Montreal, Canada.

Presented as the Presidential Address to the Psychometric Society's Annual Meeting, May, 1982. I wish to express my gratitude to my colleagues in France, especially at the University of Grenoble, for their warm hospitality during my sabbatical leave. Preparation of this paper was supported by Grant APA 0320 from the Natural Sciences and Engineering Research Council of Canada.

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

0033-3123/82/1200-5004$00.75/0

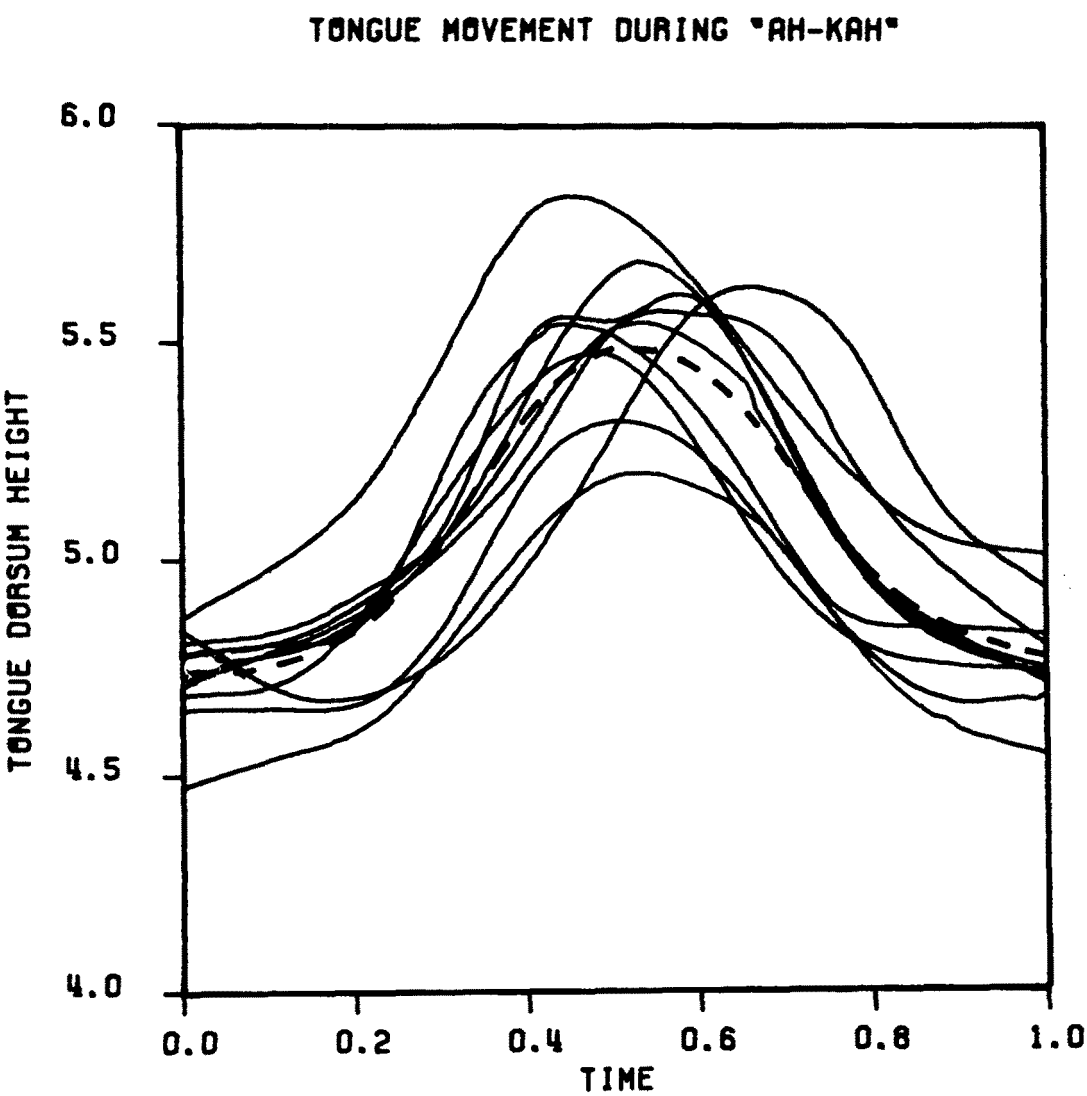


FIGURE 1
The height of the tongue dorsum over a 400 millisecond interval of time during which the sound "ah-kah" was uttered. Each curve represents a single utterance. The same subject was involved in all ten replications. The average curve is represented by a dashed line and was computed by averaging the ten curves at each point in time. The time units have been arbitrarily scaled to the interval [0, 1].

paper will be correctly perceived by many of the readers of *Psychometrika* as being a generalization of the pioneering work of Tucker [1958], and it is a privilege to again acknowledge the work of someone who has so often been there first.

Figure 2 offers an approach to the concept of a functional datum. In the upper left corner we have the domain of the classical data matrix: each of n subjects is paired with each of p variables and to each pair a number x_{ij} is assigned as the consequence of an experiment or data collection. As one moves down from this corner, we come to the situation where n is in effect infinity and we are discussing population characteristics.

Let us now fix the number of subjects n and allow the number of variables p to increase without limit. This process can be extended even beyond countability to the

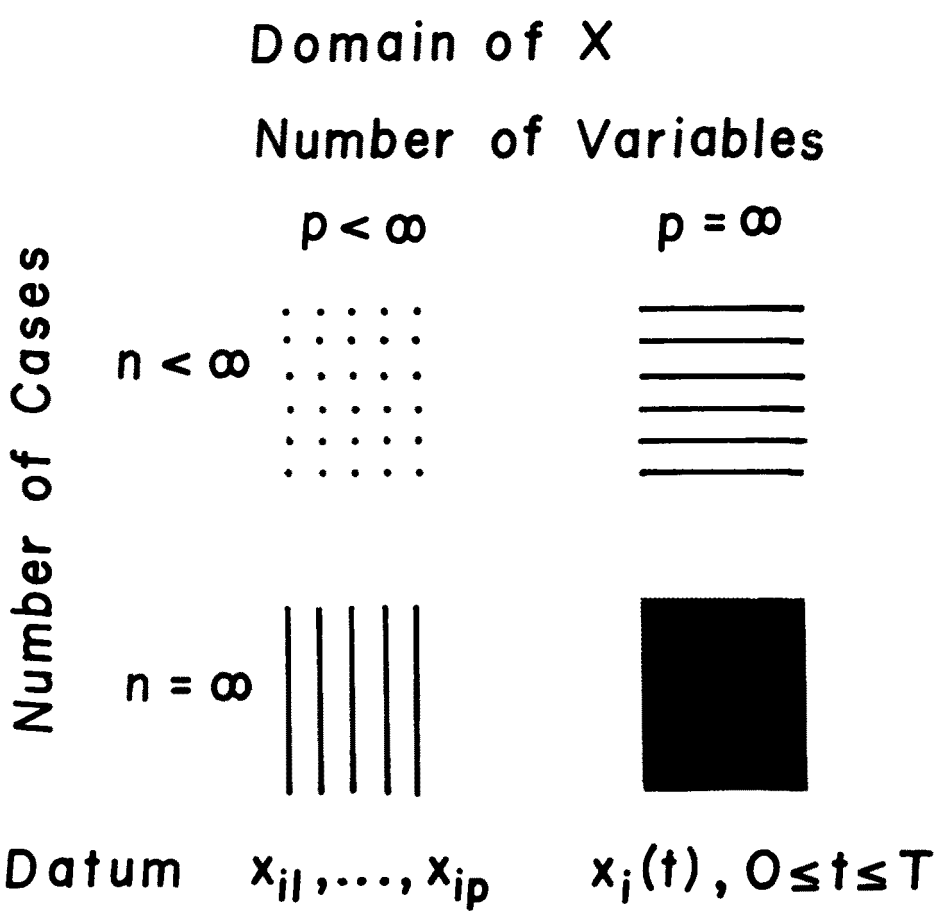


FIGURE 2
Possible domains for statistical observations. These domains depend on whether the number of replications or cases (n) is finite or infinite and whether the number of variables or points of observation (p) is finite or corresponds to the points on a continuum.

situation where the variables define a continuum. As the upper right corner of Figure 2 indicates, the data now offer a number for each point on a continuum for each subject, and it becomes natural to use function notation $x_i(t)$ to designate the value assigned to individual i at point t on this continuum. The lower right corner shows that one may even have an infinity of subjects or cases to consider, although in this talk I will confine my attention to finite n.

When presented with a continuous function, classical statisticians have tended to respond in two ways. The first is to sample the continuum at a limited number of points $t_j, j = 1, \ldots, p$. In effect each point is considered as a variable in the classical multivariate problem. There are a number of disadvantages to this approach, however. Continuity and higher orders of smoothness, which are usually important aspects of functional data, are ignored. Information between sampling points is lost. Covariance parameters proliferate with murderous speed in p so that a reduction of information does not achieve a reduction in model complexity.

The second classical response has been to postulate a family of functions which approximate the data but which depend on a limited number of parameters. The literature on the learning curve and the item characteristic curve illustrates this process nicely. Until

recently this approach ran into problems with the lack of flexibility of most parametric curve families, but the advent of spline function technology has been a great breakthrough. Nevertheless, summarizing the data in terms of a point in parameter space is not the same thing as summarizing directly as functions. This problem becomes particularly obvious when one tries to express variation in the functional data in functional rather than point terms.

The major step required to accommodate functional data is to express traditional statistical ideas in functional analytic terminology. This involves thinking of the data as defining a mapping rather than a set of points. That is, the data must be viewed as an element in a space of possible functions taking a domain space into a range space. After describing what this means first for the p-variate case and then for the functional data case, I will look at least squares estimation, principal components analysis, and canonical correlation analysis from this perspective. In each case the transition from classical multivariate data to functional data is very simple, involving essentially replacing a summation by an integral. The essential and perhaps most difficult step is the change from a static to a dynamic conceptualization of the data.

A Functional Analytic View of the Data

The best way to begin is by a reminder that a *vector space* is a set of entities which can be added according to the usual rules of addition and can also be multiplied by numbers or scalars. Scalar multiplication also distributes with respect to vector addition. The most familiar example of a vector space is the set of p-tuples of real numbers, but it is very important to think more generally. Literally anything that can be added and stretched or shrunk by an arbitrary factor can qualify as a vector. A most important alternative example is the set of vector-valued continuous functions of a set. The sum of any two functions will be a continuous function and hence also within the set. Since the functions are vector-valued they can be multiplied by a scalar, and since this preserves continuity the resulting function is still in the space. A vector space of functions particularly relevant to this paper is the space of real-valued functions having squares which have a finite integral over their domain.

There is a fundamental distinction between vector spaces which are finite dimensional and those which are not. Elements of the former can be represented as weighted sums of a finite number p of *basis vectors*, and can be put into one-to-one correspondence with p-tuples of numbers. Infinite dimensional spaces, which includes most spaces of functions, may not be representable in this way or may require an infinite number of basis vectors. Function spaces contain useful finite dimensional subspaces, however, consisting of functions which are linear combinations of a finite number of special functions. Thus, the set of polynomials of degree $p - 1$ is a p-dimensional subspace of the space of continuous functions.

An especially useful type of vector space is one which is equipped with an *inner product*. This is a real-valued function of two vectors which is symmetric and linear in its arguments, and which produces a positive number when both arguments are the same nonzero vector. The familiar inner or dot product of two p-tuples is an example. In spaces of square-integrable functions the integral of the product of two functions is an inner product.

In the following discussion of data analysis from a functional analytic point of view, the familiar p-dimensional vector space of p-tuples of real numbers will be juxtaposed

with the vector space of real-valued functions defined over a closed interval and for which the integral of their squares is finite.

p-variate Data

Let X represent the set of observations x_{ij}, $j = 1, \ldots, p$, $i = 1, \ldots, n$. X can be viewed as a funcion or a mapping from one vector space into another. The two vector spaces in question are:

1. *Subject space E of dimension p.* In this space any subject or case, observed or hypothetical, is represented by a point corresponding to its relation to each of the p variables. In this space two vectors e_j and e_k will have an inner product $b_E(e_j, e_k)$, which for purposes of simplicity will be $e_j^t e_k$, and the norm $\| e \|$ of a vector e is the square root of the inner product of e with itself. The space can be spanned by a set of p orthonormal vectors e.
2. *Variable space F of dimension n.* In this space any variable, observed or hypothetical, is represented by the values assigned to the n subjects at that position. It also has an inner product $b_F(\cdot, \cdot)$ and a set of n orthonormal spanning vectors.

The mappings that X represents are then

1. X: $E \dashrightarrow F$. Let e be any element of E. Then with X arranged as an n by p matrix the matrix product $f = Xe$ is an n-vector and thus a position in variable space.
2. X^t: $F \dashrightarrow E$. For any element $f \in F$ the product $X^t f$ is an element of E. The mapping X is called the *transpose mapping* associated with X. This has a broader meaning in functional analytic terms* for any elements e of E and f of F a mapping and its transpose satisfy $b_F(f, Xe) = b_E(X^t f, e)$.

It can be useful to think of spaces E and F as themselves spaces of functions. For example, F can be considered the space of possible mappings of n individuals into the real number line. One then has a complete expression of the problem in function space terminology. As unfamiliar as this may be, it can pay handsome dividends when considering some new problems. However, the main temptation to be resisted at this point is to consider a column or row of X as an element of F or E, respectively; X must be viewed as a mapping rather than a set of vectors.

In addition to the two mappings above associated with X there are two other important mappings determined by X. Both of these are called *operators* since they are mappings from a space into itself.

3. V: $E \dashrightarrow E$. The fact that X maps a vector from E to F and X^t maps it back again means that the composite $X^t \circ X$ maps E into itself. Denoting the matrix $X^t X$ by V, this mapping corresponds to the matrix product Ve. When the matrix X has zero column means, the matrix $n^{-1}V$ is the variance-covariance matrix, and the corresponding operator is also called the variance-covariance operator. Figure 3 shows an example for $n = 3$ and $p = 2$ of how a particular vector is transformed by V.
4. W: $F \dashrightarrow F$. Along with operator V for E goes a corresponding operator for F which results from applying X^t to an element f and then X to the resulting image of f in E. In matrix terms $W = XX^t$.

* When one or both spaces are equipped with something other than the identity metric it is important to distinguish between the transpose of a mapping and its adjoint. The appendix discusses this.

The mappings that result from the data X can be summarized neatly in the following diagram:

$$V \subset E \underset{X'}{\overset{X}{\rightleftarrows}} F \supset W. \tag{1}$$

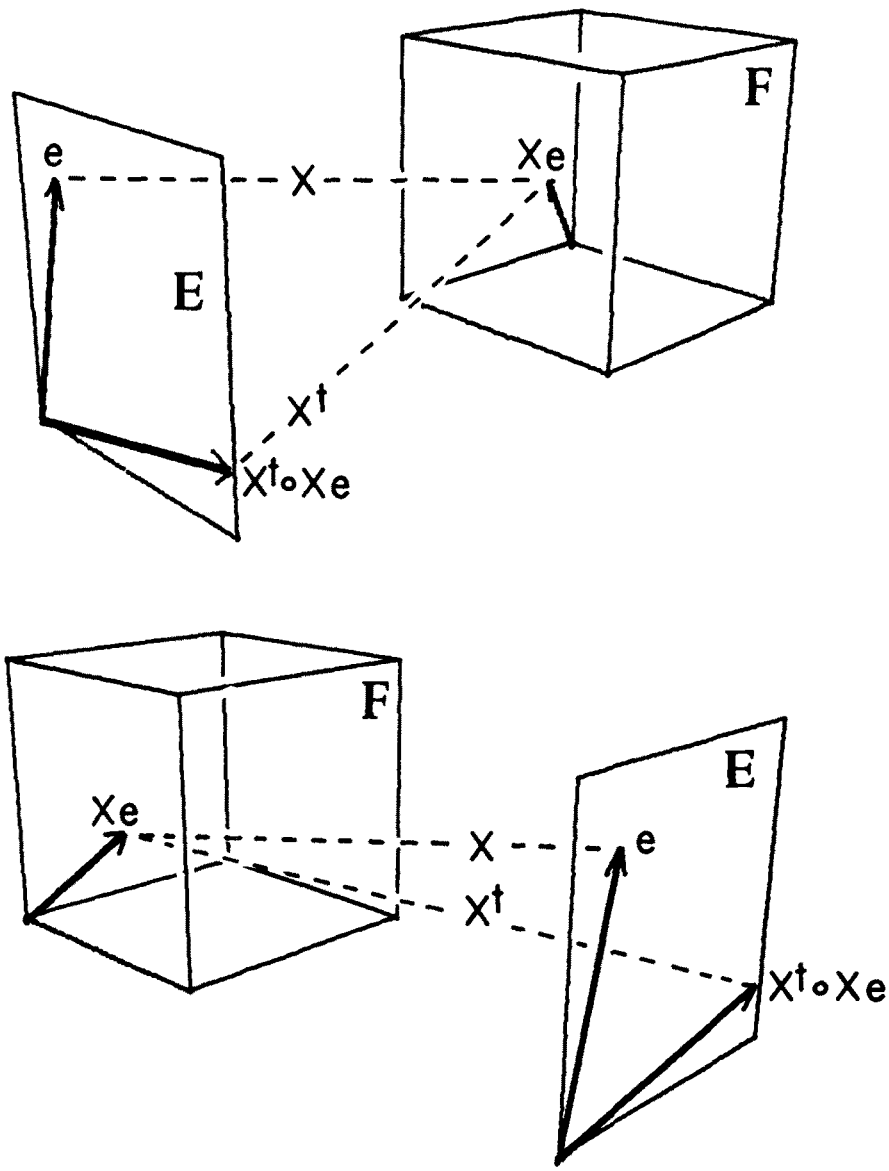


FIGURE 3
Two views of the transformation of the vector $e = (2, 5)^1$ by the operator $V = X'X$, where X is the data matrix

$$X = \begin{bmatrix} 2 & 0 \\ 1 & 0 \\ 0 & .5 \end{bmatrix}.$$

This is a simplified version of what Jean-Pierre Pagès [Pagès & Tenenhaus, Note 2] will describe for us tomorrow* as the *duality diagram.* It is a very important symbolic tool for the presentation of data analytic problems in functional analytic terms. In this talk I will ignore any considerations involving metrics for spaces E and F. Appendix A explains how the duality diagram is extended to accommodate metrics other than the identity by the use of the dual spaces for E and F.

We may now say that the data determine a subspace of F in the following way. Any vector e in E is mapped by X into F. If we choose the p orthonormal vectors e_j in E and map them into F, the result will be p vectors in F; and the image of any e in F can be represented by a linear combination of these p images. Thus the subspace of F is in effect the image of E under the mapping X, and will usually be of dimension min $\{p, n\}$. If an arbitrary element f within F does not lie within this subspace, its image under operator W certainly will since W is a result of first mapping f into E and then its image back into F.

Functional Data

The next task is to extend the functional analytic discussion of the data given above to the situation in which the datum for individual or case i is a function $x_i(t)$, $0 \leq t \leq T$. As Figure 2 suggests, one may imagine this arising from allowing the number of variables to become so large that the index j, $j = 1, \ldots, p$, can be treated as a continuum and is renamed t. The continuum will be referred to as time in this paper since this is so often the case in practice. In general, where summation over j takes place in the classical case, integration over t will now be required. The data will still be indicated by X but it will no longer be worthwhile thinking of X as a matrix. Nevertheless, X defines mappings among the following spaces.

1. *Subject space E of infinite dimension.* Any subject, observed or hypothetical, can be associated with a function $e(t)$. Since an infinite variety of functions is possible the space has infinite dimensionality. An inner product $b_E(e_j, e_k)$ for functions $e_j(t)$ and $e_k(t)$ is given by

$$b_E(e_j, e_k) = \int_0^T e_j(t)e_k(t)\, dt,$$

   and again the norm of a function is the inner product of the function with itself; that is, the integral of its square. If attention is confined to functions having finite norms, then the resulting space is a *Hilbert space.* It is fundamental to Hilbert spaces that any element can be represented in them by a weighted sum of a countable number of orthonormal functions, so that although the dimensionality is infinite, it is at least countable. E in the p-variate case is also a Hilbert space.
2. *Time space F of dimensionality n.* This space has exactly the same characteristics as F in the p-variate case. Any point in time, observed or hypothetical, is represented in this space by the values of the n functions at that time point.

The mappings that X represents are now

1. $X: E \dashrightarrow F$. Let $e(t)$ be any function in space E. Then the vector f whose ith element is given by

$$f_i = \int_0^T x_i(t)e(t)\, dt \tag{2}$$

   is an element in n-space and hence in F. Note that this is formally equivalent to

* (See Footnote 1 p. 379)

what one would obtain in the p-variate case if the summation for the ith element in the product Xe were replaced by an integration. Thus, it may be helpful to imagine X in the functional data case as represented by a matrix with completely dense rows.

2. X^t: $F \dashrightarrow E$. Let f be any vector in F. Then the function

$$X^t: f \dashrightarrow X^t f = e(t) = \sum_i^n f_i x_i(t) \tag{3}$$

is an element of E. Moreover, it is easy to see that $b_F(Xe, f) = b_E(e, X^t f)$ for any $e \in E$ and $f \in F$ so that one is justified in using the notation X^t for this mapping.

3. V: E $\dashrightarrow E$. Again the consequence of mapping a function $e(t)$ into F by (2) and then mapping the resulting n-vector back into E by (3) can be described as an operator $V = X^t \circ X$. It is analagous to a symmetric matrix with order so large that it appears to have completely dense rows and columns. Explicitly

$$X^t \circ X: e(t) \dashrightarrow Ve(t) = \sum_i^n x_i(t)\left[\int_0^T x_i(u)e(u)\,du\right]$$

$$= \int_0^T \left[\sum_i^n x_i(t)x_i(u)\right]e(u)\,du. \tag{4}$$

Note that V is a member of that general class of integral transforms representable by $\int K(t, u)e(u)\,du$, where the function $K(t, u)$ is called the *kernel* of the transform. In this case $K(t, u) = \sum x_i(t)x_i(u)$.

4. W: $F \dashrightarrow F$. The consequence of mapping a vector f in F into E and back again is a vector whose ith element is

$$f_i \overset{W}{\dashrightarrow} \int_0^T x_i(t) \sum_l^n f_l x_l(t)\,dt. \tag{5}$$

These mappings are still described schematically by the duality diagram (1). The image of E in F determined by the transformation X is now in general of dimension n and thus coincides with F itself. This does not mean, however, that a W maps a vector f into itself, and a central problem is now to study the consequences of any or all of these mappings.

Least Squares Approximation in Functional Terms

The key idea underlying the most commonly used classical statistical procedures is the approximation of a set of points by points lying within a subspace of reduced dimensionality. For example, multiple regression is the process of representing an element f in F by its image $\hat{f}$ in the k-dimensional subspace spanned by predictor variables f_j, $j = 1, \ldots, k$. This image $\hat{f}$ is a result of a *projection* of f onto this subspace. Thus, we may consider $\hat{f}$ to be the consequence of applying a projection operator P to f, where P maps any vector into its least squares image in the subspace. Projection operators have two important properties: $P \circ P = P$ and $P^t = P$. Implicit in the use of the term "operator" is the uniqueness of the least squares estimate, which is the case when the vector being approximated is (a) an element in a Hilbert space, and (b) the approximation is a member of either a closed subspace or a convex subset.

Approximation problems in classical p-variate data analysis are usually expressed in terms of the variable matrix F. However, it can be very useful in the functional data

case to consider the problem of approximating a function $e(t)$ in E by its projection on a finite-dimensional subspace E. This process is familiar to electrical engineers who apply filters to input signals to eliminate unwanted components. This is also what the numerical analyst does in approximating a complicated function in terms of a linear combination of simpler functions. When the goal of the approximation is the minimization of $\| e(t) - \hat{e}(t) \|^2 = \int [e(t) - \hat{e}(t)]^2 \, dt$ the corresponding mapping is a projection P. This can be represented by the duality diagram

$$\hat{E} \underset{P}{\overset{P}{\rightleftarrows}} E. \tag{6}$$

One set of approximating functions which have spectacular properties are piece-wise polynomials or splines, which are not only very flexible with only a modest number of parameters but remarkably easy to handle computationally. Winsberg and myself [1980; 1981; Winsberg & Ramsay, Note 3] have been working with monotone splines, and she will report on another application in these meetings [Winsberg & Ramsay, 1982, Note 4]. Monotone spline approximation involves projecting function space onto a cone and has turned out to be not at all difficult in a wide variety of problems. A comprehensive treatment of splines and the various spaces associated with them is to be found in Schumaker [1981].

Principal Components Analysis

This very important technique is usually motivated in classical statistics by the problem of either approximating the variance-covariance matrix $n^{-1}X'X$ of rank p by a matrix of reduced rank k, or of approximating the data matrix X by a matrix $\hat{X}$ of the same dimensions but reduced rank k. In either case the solution can be expressed in terms of either the eigenanalysis of $X'X$ or the singular value decomposition of X. Counterparts of eigenanalysis or the singular value decomposition for Hilbert spaces of functions also play the critical role in the principal components analysis of functional data.

From a functional analytic point of view the problem is one of the description of a mapping $V: E \dashrightarrow E$ or alternatively of $X: E \dashrightarrow F$. Thus, we seek mappings $\hat{V}$ or $\hat{X}$ which are in some sense close to those they approximate. One way in which the concept "close" can be expressed is by saying that the consequences of the approximating mapping should be as close as possible to those of the original. In this way, the problem of approximating a mapping can be reduced to that of approximation in the range space of the mapping, and hence can be put into the least squares terminology discussed in the previous section.

The consequences of mapping X can be summarized by describing what happens to a vector e in E of unit norm when it is mapped into F, whereupon it has norm $\| Xe \|$. Now the image of the unit hypersphere in E will be a hyperellipsoid in F because the mapping of X is linear. Figure 4 displays the image of the unit circle resulting from the 3 by 2 data matrix displayed in Figure 3. Thus, the location of an element e_i in E whose image $\| Xe_1 \|$ in F has the largest norm can be seen as a best one-dimensional approximation to this hyperellipsoid. The location of an element e_2 orthogonal to e_1 and having the image with the largest norm provides a best two-dimensional approximation. This process can be continued until either the patience of the data analyst or the dimensionality of E is exhausted.

Because of the way in which the transpose was defined above,

$$b_F(Xe, Xe) = b_E(e, X^t \circ Xe) = b_E(e, Ve). \tag{7}$$

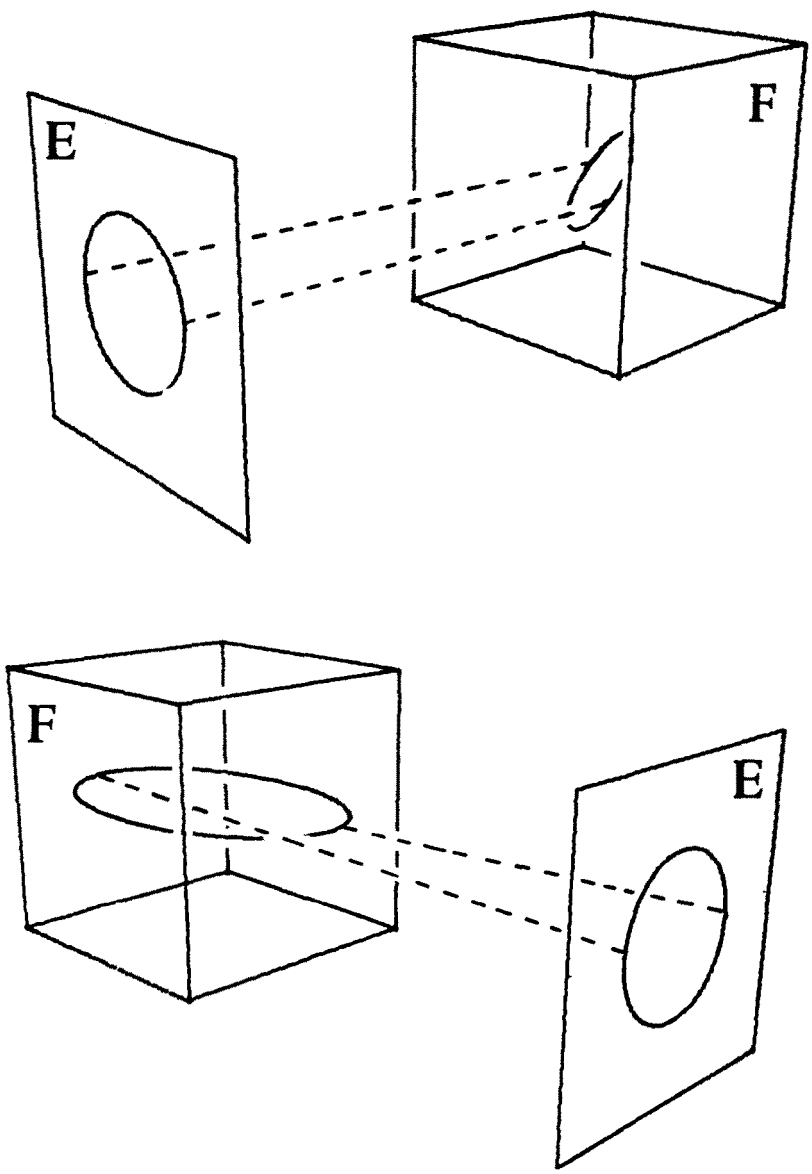


FIGURE 4
Two views of the transformation of the unit circle by the mapping of X, where X is the data matrix

$$X = \begin{bmatrix} 2 & 0 \\ 1 & 0 \\ 0 & .5 \end{bmatrix}.$$

Thus, the search for vector e of unit norm yielding maximum $\| Xe \|$ is equivalent to the search for the element maximizing $b_E(e, Ve)$. Any inner product satisfies the Cauchy-Schwarz inequality, $b_E(e_j, e_k)^2 \leq b_E(e_j, e_j)b_E(e_k, e_k)$, and it follows from this that $b_E(e, Ve)$ is maximized when

$$Ve = \lambda e. \tag{8}$$

This is of course an *eigenequation.* From (7) we have that $\| Xe \|^2 = \lambda$ and thus that λ is a measure of variation in F. One may consider Xe as defining the dominant direction of variation in the image of E in F, or in the image of F under mapping W. Under fairly general conditions it can be shown that operator V can be expressed as $\sum \lambda_j e_j(u) e_j(t)$ with at most a countable number of terms and where (λ_j, e_j) is the jth solution to (8).

The identification of solutions to this equation is called *spectral analysis,* and is one of the central topics in functional analysis. In the finite p case e is an eigenvector of symmetric positive definite matrix V. In the functional data case $e(t)$ is an eigenfunction of the operator V. In either case λ is an eigenvalue of V. Under very general conditions which are reasonable for practical work it can be shown that the number of solutions to the eigenequation are at most countable, that the eigenvalues are nonnegative and distinct, and that the largest eigenvalue is finite.

There are various ways of tackling the problem of computing the eigenfunctions $e_j(t)$. When n is not too large one may perform the matrix eigenanalysis of the matrix W instead. The required eigenfunctions are then simply $X^t f_i$, $i = 1, \ldots, n$. Techniques for dealing directly with the operator V also exist, some of which involve discrete approximations for which an adequate convergence theory exists. A simple form of discrete approximation to operator V involves choosing a sufficiently large number of equally spaced points in $[0, T]$ and approximating the integrals in (4) by the corresponding sums divided by the number of points. In this form the problem becomes identical with the classical multivariate procedure for principal components analysis. However, more sophisticated quadrature procedures will require much fewer points in general.

An important extension of principal components analysis is when E is mapped into yet another space G according to the following diagram:

$$G \underset{U^t}{\overset{U}{\rightleftarrows}} E \underset{X^t}{\overset{X}{\rightleftarrows}} F. \tag{9}$$

A particularly important example arises when G is a subspace or a convex subset of E. For example, G may consist of the space spanned by a set of spline functions, or the cone consisting of convex combinations of monotone splines. In this case both U and U^t will be the projection operator P. What is essential is that one may again define principal components analysis as the extremal problem max $\| X \circ Ug \|$ subject to $\| g \| = 1$.

Descriptive Statistics for Functional Data

It is now possible to discuss the problem of descriptive statistics for functions in more detail. The location of a set of functions can be summarized by the point-wise average over subjects. Alternatively, if the functions have a large noise component or other undesirable nonsmooth components, it may be preferable to first approximate them by suitable splines and then take the average of the approximations. The original functions can then be centered by subtracting the average function from each of them.

Summarizing dispersion is more complicated. Just as one cannot summarize the spread of a p-variate distribution in terms of the spread for each variable alone, so it is that measuring the spread of the functions point-wise will be of little value. Instead, p-variate spread is captured in the variance-covariance matrix $n^{-1}X^tX$. Analogously, it is the variance-covariance operator $n^{-1}V$ in the functional data case which contains the essential information on the dispersion of functions. The kernel of this operator $K(u, t) = n^{-1} \sum x_i(t) x_i(u)$ defines a *variance-covariance surface* and it can be highly instructive to plot this surface using either contour plotting or perspective plotting techniques. The

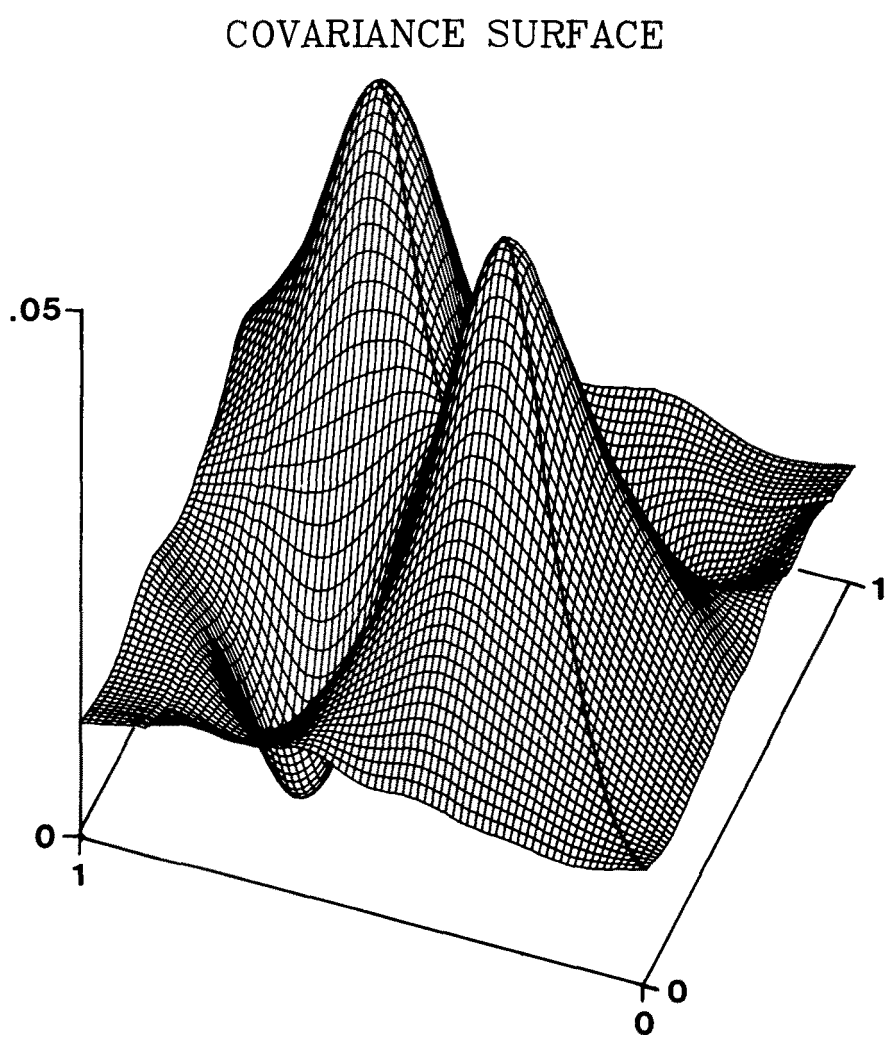


FIGURE 5
The kernel of the variance-covariance operator $K(t, u) = x_i(t)x_i(u)$ computed from the functions in Figure 1 plotted as a surface. The peaks correspond to times of about 0.4 and 0.8 when the tongue dorsum was decelerating. The hollows occur at the points (0.4, 0.8) and (0.8, 0.4) and indicate that covariance between points where the tongue dorsum is decelerating is low. The height of the surface along the diagonal is the variance function.

correlation surface defined by $K(u, t)/[K(u, u)K(t, t)]^{1/2}$ may also be plotted. The eigenfunctions $e_j(t)$ corresponding to dominant eigenvalues indicate the dominant types of variation from the average function, just as the eigenvectors in p-variate analysis indicate the dominant directions of variation. Thus they should also be displayed.

The family of techniques known as time series analysis can be linked to the analysis of functional data at this point. These procedures are based on the concept of *stationarity* of covariance structure which implies that the correlation surface $K(u, t)$ can be defined as a function of $|u - t|$ alone [Doob, 1953]. It can be shown that the eigenfunctions in this case are periodic and hence representable by at most a countable combination of sines and cosines.

The descriptive analysis of functional data can be illustrated by the analysis of the

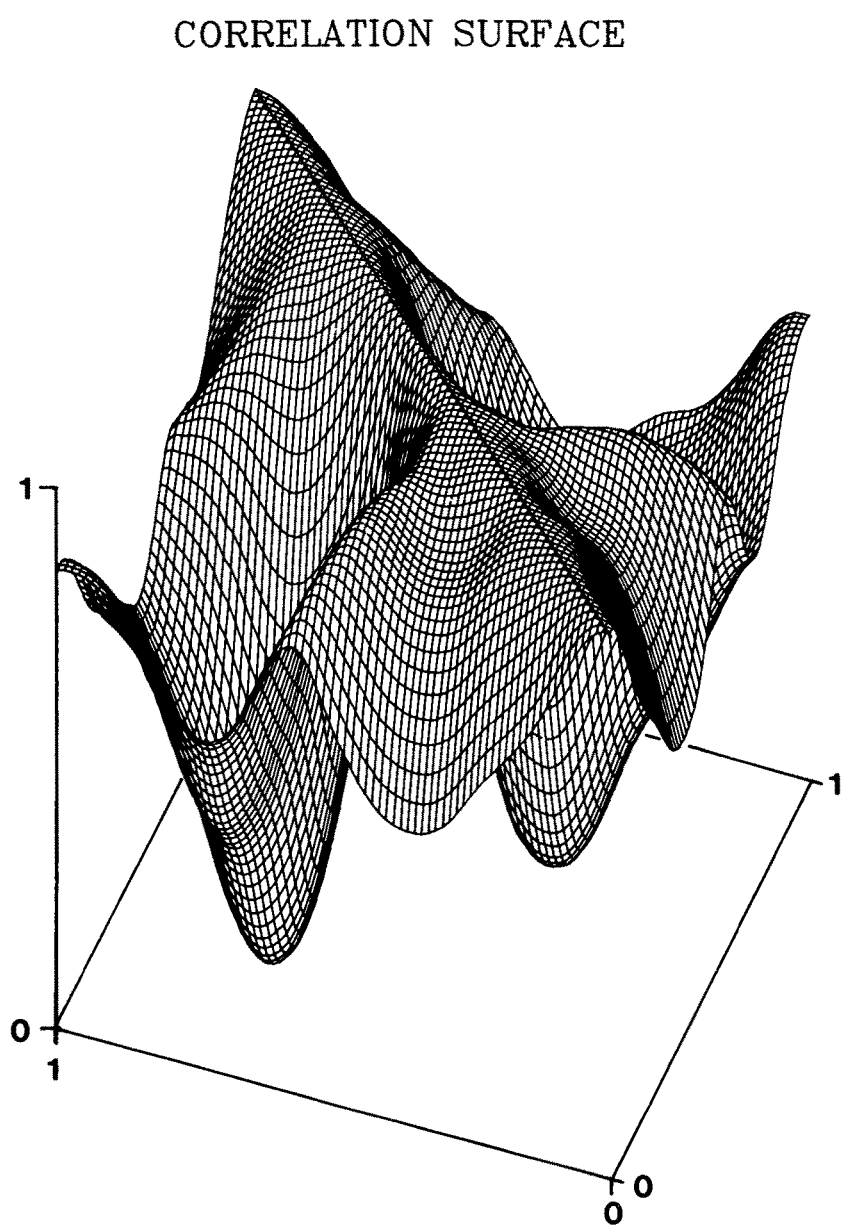


FIGURE 6
The correlation surface $K(t, u)/(K(t, t)K(u, u))^{1/2}$ computed from the function in Figure 1. The hollows at (0.4, 0.8) and (0.8, 0.4) indicate that tongue dorsum movement is uncorrelated with itself at points where the dorsum is decelerating.

curves presented in Figure 1, in which the pointwise average curve is indicated by a dashed line. Note that the tongue dorsum is strongly decelerating at $t = 0.4$ and $t = 0.8$ and reaches its maximum height on the average at $t = 0.5$. Figure 5 displays the variance-covariance surface for these curves. The height of the surface along its diagonal running from lower right to upper left indicates the variability of the curves at each point in time. The twin peaks correspond to $t = 0.4$ and $t = 0.8$ when the tongue dorsum is decelerating. The correlation surface is displayed in Figure 6 and again the deep wells in the surface

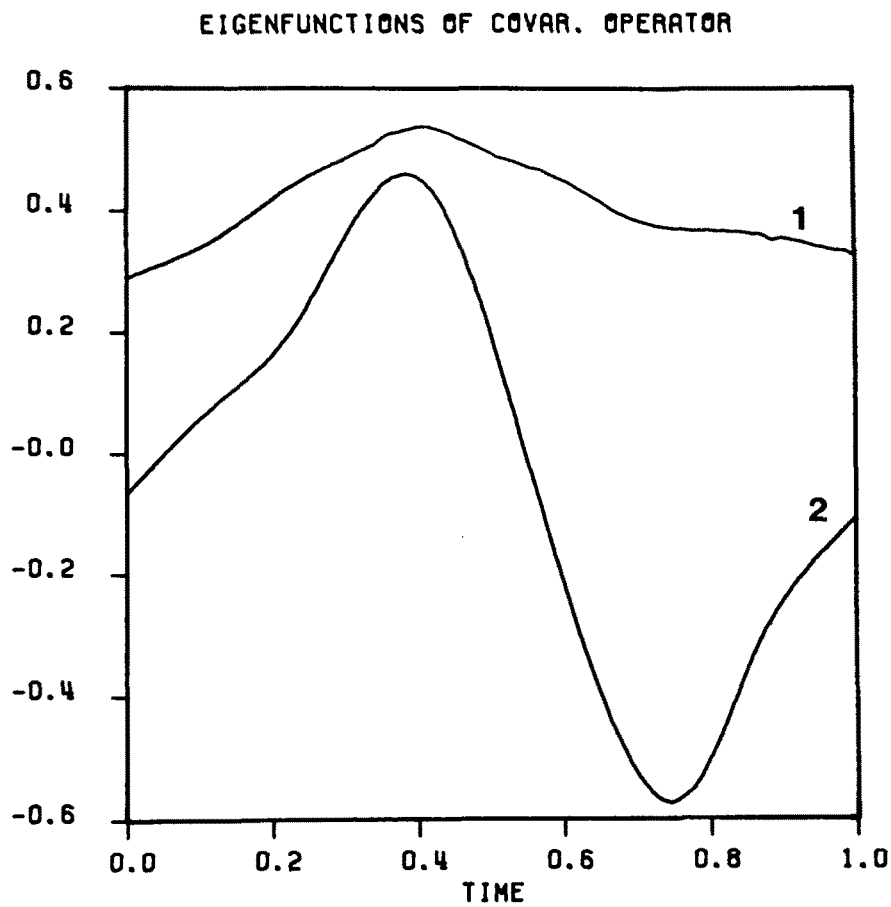


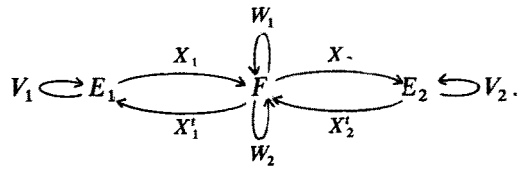
FIGURE 7
The first two eigenfunctions of the variance-covariance operator $K(t, u)$ computed from the data in Figure 1. These correspond to eigenvalues of 0.17 and 0.12, respectively and account for 95% of the total of all eigenvalues. The first eigenfunction indicates that the dominant form of deviation from the mean function is a vertical shift, while the second eigenfunction indicates that a second component of variation is under- or over-shooting of the tongue dorsum at points where it is decelerating.

correspond to the near zero correlation between tongue dorsum height at the two points of deceleration. It appears that the tongue starts up with precision but has some difficulty slowing down. The first two eigenvalues account for 95% of the total, and Figure 7 displays the first two eigenfunctions. These may be summarized by saying that the two dominant modes of variation are a general vertical shift of the function, and an initial overshooting prior to maximum height followed by an undershooting at minimal height.

Canonical Correlation Analysis

The description of canonical correlation analysis in functional analytic terms has a great deal to offer a number of problems, especially since so many familiar techniques in multivariate data analysis can be embedded within it. Let us now suppose that we have two subject spaces E_1 and E_2. In each a subject is located in terms of his relation to a set of variables or points in time, with two different sets of variables or time being involved. Let the two data-determined mappings from E_1 and E_2 into F be denoted by X_1 and X_2, respectively. Then there are two subspaces in F to be considered. The first F_1 is the image

of F under the operator $W_1 = X_1 \circ X_1^t$ and the second F_2 its image under $W_2 = X_2 \circ X_2^t$. Corresponding to these two subspaces are the projections P_1 and P_2 taking any element of f into its least squares approximation in each of the subspaces F_1 and F_2, respectively. This situation can be summed up in the following simplified duality diagram:



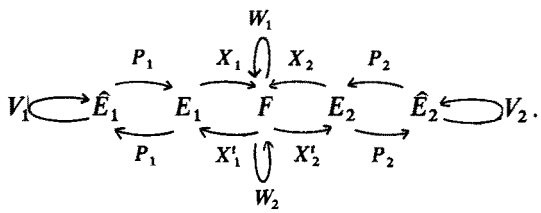
(10)

Canonical analysis can be expressed as the description of $F_1 \cap F_2$. In order for the analysis to be nontrivial one must impose the additional requirement that these subspaces be closed, which in practice means finite dimensional. An arbitrary element f in F will be mapped into this intersection space if it is first projected into F_1 by projection P_1 and then this image is projected into F_2 by projection P_2. This means that the operator $P_1 \circ P_2$ (or for that matter $P_2 \circ P_1$)provides the required mapping.

In general terms canonical analysis reduces to the spectral analysis of the operator $P_1 \circ P_2$. In the case where E_1 and E_2 are of finite dimensions p and q, respectively, this reduces to the eigenanalysis of $(X_1 V_1^{-1} X_1^t)(X_2 V_2^{-1} X_2^t)$.

When one expresses the canonical correlation problem as one involving the spectral analysis of the product of operators, it is obvious that it may be generalized to any number of sets of variables or functions. Thus canonical correlation analysis for k-way tables involves the eigenanalysis of the operator $P_1 \circ P_2 \circ \cdots \circ P_k$.

When the data for each subject is a pair of functions, the projections involved will not normally be closed subspaces of F. In practical terms this means that two sets of functions may correlate highly in terms of arbitrarily small and kinky components. The way around this difficulty is to project each function space E onto an appropriate subspace or convex subset $\hat{E}$ of finite dimensionality (and less than n). This may be viewed as a process of filtering the data. Using the fact that a projection is its own transpose, this results in the following duality diagram:



(11)

Now it is in general the case that the operators $V_k = P_k \circ X_k \circ X_k^t \circ P_k$, $k = 1, 2$, will have an inverse and the analysis may proceed as above.

In the functional data case it is natural to consider interchanging the roles of E and F in the duality diagram. When the subjects are partitioned into two or more groups one may be interested in determining pairs of functions, each being a weighted sum of functions within a particular group, which have maximal intercorrelation subject to the usual orthogonality conditions. Although such an analysis would seldom be of much interest in the p-variate case, where the ordering of variables is usually arbitrary, the ordering of points of time could make such an analysis very useful.

Conclusion

I have tried to indicate in a very general way how statistical concepts can be extended to include functional data. This requires the point of view from which all data are

functional, and finally there is no essential distinction between data domains which are finite sets and those which are continua. Functional analytic terminology also permits a natural and rigorous treatment of a number of other familiar and not so familiar situations, including dual scaling or correspondence analysis, vector-valued continuous functions as data, and tables of data with an arbitrary number of modes. At least one branch of statistics is already expressed in functional terms: Bayesian inference is essentially the mapping of the space of density functions defined on a parameter space into itself using a nonlinear operator determined by the data. In this sense operator V is only a special case.

Unfortunately we in North America are handicapped by the fact that a course in functional analysis is seldom a part of the preparation of an applied statistician. As was the case for linear algebra, we pragmatic souls demand a demonstration of practical relevance before making such a commitment. There are now a number of elementary introductions to functional analysis, including Aubin [1979] and Kreyszig [1978]. However the classic work by Dieudonné [1960] still has few rivals. A very elementary treatment of this material in a statistical context for the finite p case is a wonderful book by Cailliez and Pagès [1976] which is, alas, still only in French.

I would like to conclude with a quote from Dieudonné [1960].

> The student should as soon as possible become familiar with the idea that a function f is a single object, which may itself "vary" and is in general to be thought of as a "point" in a large "functional space"; indeed, it may be said that one of the main differences between the classical and the modern concepts of analysis is that, in classical mathematics, when one writes $f(x)$ f is visualized as "fixed" and x as "variable," whereas nowadays *both* f and x are considered as "variables" ... (p. 1)

Functional analysis already has revolutionized numerical analysis, so that any issue of a major journal now has a number of papers using this technology. I claim that this is about to happen in statistics, and I hope that my talk leads you to speculate on this possibility.

Appendix

The inner product function used in the text is too simple for many applications. For example, the use of standardized data is equivalent to the use of inner product

$$b_E(e_j, e_k) = \sum_m^p \frac{e_{mj} e_{mk}}{\sigma_m^2}.$$

Alternatively, one can say that $b_E(e_j, e_k) = e_j^t M e_k$, where M is a diagonal matrix containing reciprocals of variances in its diagonal. Matrix M is called the *metric* for E, and a more general treatment requires the consideration of metrics M and N for spaces E and F, respectively. The use of metrics requires the concept of a *dual space*, and in this appendix the duality diagram (1) is extended to include these dual spaces.

For any vector e there are various possible real-valued functions that can be computed. Weighted sums of elements and the maximal element are two examples. The set of possible continuous real-valued functions of e which are linear in e is itself a vector space since it is closed under weighted summation and has the other vector space properties. This space is denoted by E^* and is referred to as the dual of E. For example, if E consists of the space of column vectors with p elements, then the dual of E is the space of row vectors of size p. It is usual to use the notation $\langle e, e^* \rangle$ for the real number which results when a function e^* in E^* is applied to an element e of E. For example, if e^* is a row vector and e a column vector then the matrix product e^*e is denoted by $\langle e, e^* \rangle$. The symmetry of the notation corresponds to the symmetry of the relation between E and E^* since the dual of E^* turns out to be E.

There is a one-to-one correspondence between the two spaces induced by the equation

$$b_E(e, e_0) = \langle e_0, e^* \rangle, \tag{A1}$$

where e_0 is any fixed nonzero element of E. This equation defines a mapping $M: E \dashrightarrow E$ for which to any $e \in E$ is associated the element $e^* = Me$ in E which satisfies (1). In the p-dimensional case both E and E^* can be represented by p-tuples of numbers, and $e^* = Me$, where M is the metric matrix. Thus M is used to denote both a metric and the mapping from E to E^*. If E^* is equipped with inner product $b_{E^*}(e_j^*, e_k^*) = e_j^{*t} M^{-1} e_k^*$ and thus with metric M^{-1}, then the two inner products satisfy $b_{E^*}(e_j^*, e_k^*) = b_{E^*}(Me_j, Me_k) = b_E(e_j, e_k)$.

It is usual to postulate two further properties of E. One of these is *completeness*, implying that convergent sequences of vectors converge to a vector in the space. Intuitively this implies that the space has no "holes" in it; the real numbers are a complete set since they have no gaps between them while the rationals are not because there are sequences of rationals converging to irrational numbers. A complete vector space with an inner product is called a *Hilbert space*. The other property used in practice is *separability*. This says that the space contains a dense countable subset: every neighborhood of any element contains a member of this dense countable subset. For example, every neighborhood of a real number contains a rational number. Intuitively, this condition implies that the space is "evenly and thinly spread."

Consider now a continuous linear mapping $T: E \dashrightarrow F$ from Hilbert space E into another Hilbert space F. Then there are two other mappings automatically associated with T:

1. $T^*: F \dashrightarrow E$: the *adjoint* mapping defined by the following equation

$$b_F(f, Te) = b_E(T^*f, e) \tag{A2}$$

for any $e \in E$ and $f \in F$.

2. $T^t: F^* \dashrightarrow E^*$: the *transpose* mapping defined by the following equation

$$b_F(f, Te) = b_{E^*}(T^t f^*, e^*), \tag{A3}$$

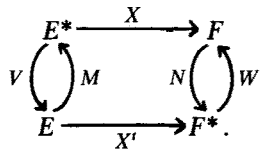
where $f^* = Nf$ and $e^* = Me$.

In the special case where both E and F are equipped with identity metrics, the two mappings are equivalent. The transpose mapping has the advantage of being invariant with respect to changes in metric. In the finite dimensional case with E equipped with metric matrix M, F with metric matrix N, and the mapping T represented by matrix T, the transpose mapping is represented by T^t and the adjoint by $M^{-1}T^tN$.

Turning now to the data analytic applicaton of this theory, the data matrix X logically represents a mapping from E^* to F. This arises because of the following two considerations:

1. The ith individual is represented by a vector e_i in E specifying his relationship with the p variables or with each point in time. If the dual space E^* is spanned by a set of orthonormal basis vectors $e_j^*, j = 1, \ldots, p$, then $\langle e_i, e_j^* \rangle = x_{ij}$, the score of individual i on variable j. Thus associated with each variable j is the orthonormal basis vector e_j^*.
2. The jth variable is also represented by vector f_j in F specifying the relationship of this variable with the n individuals. Again, if f_i^*, $i = 1, \ldots, n$, is the orthonormal spanning system of vectors for F^* then $\langle f_j, f_i^* \rangle = x_{ij}$, and associated with each individual i is the orthonormal basis vector f_i^*.

Thus in both E^* and F are to be found vectors corresponding to any variable, and $X: E^* \dashrightarrow F$ maps e_j^* into f_j. Similarly $X^t: F^* \dashrightarrow E$ associates individual vectors. These relations are summed up in the complete version of the duality diagram:



(A4)

The simplified version of the duality diagram given in (1) was possible because, when M and N are the identity mappings, each space and its dual have the same inner product and metric and are thus in every way identical. The mapping $V: E^* \dashrightarrow E$ is constructed by going around the diagram the long way; $V = X^t \circ N \circ X$. Similarly $W = X \circ M \circ X^t$. The principal components analysis problem reduces then to the eigenanalysis of $V \circ M: E \dashrightarrow E$ represented in the finite dimensional case by the matrix X^tNXM.

Aside from the ease with which the duality diagram permits one to extend data analysis to functional data, it also describes techniques such as principal components analysis and canonical correlation analysis in a manner that is free of both basis and metric for either E or F. In this sense it can be considered a fundamental algebraic advance over matrix analysis, and it is to be hoped that it will become a standard part of statistical language.

REFERENCE NOTES

1. Keller, E. & Ostry, D. J. Computerized measurement of tongue dorsum movements with pulsed echo ultrasound(*a*). Manuscript submitted for publication to *Journal of the Acoustical Society of America*, 1982.
2. Pagès, J. P. & Tenenhaus, M. Geometry and duality diagram. An example of application: The analysis of qualitative variables. Paper presented at the Psychometric Society Annual Meeting, Montreal, Canada, 1982.
3. Winsberg, S. & Ramsay, J. O. Monotone spline transformations for dimension reduction. Submitted for publication in *Psychometrika*.
4. Winsberg, S. & Ramsay, J. O. Monotone spline transformations for ordered categorical data. Paper presented at the Psychometric Society Annual Meeting, Montreal, Canada, 1982.

REFERENCES

Aubin, J.-P. *Applied Functional Analysis.* New York: Wiley, 1979.

Cailliez, F. & Pagès, J.-P. *Introduction à l'Analyse des Données.* Paris: Société de Mathématiques Appliquées et de Sciences Humaines, 9 rue Duban, 75016 Paris, 1976.

Dauxois, J. & Pousse, A. Les analyses factorielles en calcul des probabilités et en statistique: Essai d'étude synthètique. Thèse d'état, l'Université Paul-Sabatier de Toulouse, France, 1976.

Dieudonné, J. *Foundations of Modern Analysis.* New York: Academic Press, 1960.

Doob, J. L. *Stochastic Processes.* New York: Wiley, 1953.

Kreyszig, E. *Introductory Functional Analysis with Applications.* New York: Wiley, 1978.

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

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

Winsberg, S. & Ramsay, J. O. Monotonic transformations to additivity using splines. *Biometrika, 67,* 1980, 669–674.

Winsberg, S. & Ramsay, J. O. Analysis of pairwise preference data using integrated *B*-splines. *Psychometrika, 46,* 1981, 171–186.

Final version received 7/12/82