J. R. Statist. Soc. B (1996)
58, *No.* 3, *pp.* 495–508

Principal Differential Analysis: Data Reduction by Differential Operators

By J. O. RAMSAY†

McGill University, Montreal, Canada

[Received November 1994. Final revision June 1995]

SUMMARY

Functional data are observations that are either themselves functions or are naturally representable as functions. When these functions can be considered smooth, it is natural to use their derivatives in exploring their variation. Principal differential analysis (PDA) identifies a linear differential operator $L = w_0 I + w_1 D + \ldots + w_{m-1} D^{m-1} + D^m$ that comes as close as possible to annihilating a sample of functions. Convenient procedures for estimating the m weighting functions w_j are developed. The estimated differential operator L is analogous to the projection operator used as the data annihilator in principal components analysis and thus can be viewed as a type of data reduction or exploration tool. The corresponding linear differential equation may also have a useful substantive interpretation. Modelling and regularization features can also be incorporated into PDA.

Keywords: DIFFERENTIAL EQUATIONS; FUNCTIONAL DATA ANALYSIS; PRINCIPAL COMPONENTS ANALYSIS; SPLINE SMOOTHING

1. INTRODUCTION

The data treated in this paper are functional in that each observation y_i, $i = 1, \ldots, N$, is a real function, assumed for simplicity of exposition to be defined on [0, 1]. A functional datum may also be a vector of values that can be represented effectively by a function after a suitable smoothing or interpolation procedure, and then y_i will be taken as this representing function. Data reduction for data of this nature has been termed functional data analysis by Ramsay and Dalzell (1991).

Data reduction is the simplification of the variation in the data by first identifying and then removing important components of variation, i.e. we have available a data vector y, with elements y_i, these elements being in turn points in some vector space, so that they may be, for example, scalars, real vectors of dimension p or functions. We wish to study this data vector y by separating it into two components, $y = u + e$, where *structural component* u has some simple or interpretable or otherwise interesting low dimensional structure and *residual component* e is viewed as noise, unwanted variation or possibly as data ready for further exploration.

The first identification phase in this process is optional, since the structural components may be known in advance. In the classical linear modelling of a data vector y, a design array $\mathbf{X}$ is available, and the influence of components of variation represented by the columns of $\mathbf{X}$ are removed in ordinary least squares analysis by applying the projection operator $\mathbf{Q} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

†*Address for correspondence*: Department of Psychology, McGill University, 1205 Dr Penfield Avenue, Montreal, Quebec, H3A 1B1, Canada.
E-mail: ramsay@psych.mcgill.ca

 0035–9246/96/58495

However, the identification of important components of variation may be essential before these are removed. In the principal components analysis (PCA) of a data matrix $\mathbf{Y}$, for example, the singular value decomposition $\mathbf{Y} = \mathbf{USV}^{\mathrm{T}}$ is used to identify a number m of strong singular values in $\mathbf{S}$ and the leading m columns $[\mathbf{U}]_m$ are in effect used as the carrier $\mathbf{X}$ in the above projector $\mathbf{Q}$, so that $\mathbf{Q} = \mathbf{I} - [\mathbf{U}]_m[\mathbf{U}]_m^{\mathrm{T}}$. Projector $\mathbf{Q}$ is also defined by the leading m columns $[\mathbf{V}]_m$ of the orthonormal matrix $\mathbf{V}$, in that $\mathbf{Y}[\mathbf{V}]_m$ is the image of $\mathbf{Y}$ within the m-dimensional subspace of the principal components of variation with respect to the principal axis co-ordinate system. Thus, these m vectors in $[\mathbf{V}]_m$ can be thought of as the parameters defining the projector. For functional data, the m columns of matrix $[\mathbf{V}]_m$ are replaced by the first m eigenfunctions v_j of the variance–covariance kernel.

An alternative linear transform family that may also be applied to functional data is that of linear differential operators. We shall be exploring the linear differential operator analogue of PCA, which may be called *principal differential analysis* (PDA). A mapping of this type, denoted by L, is defined as follows. Let D^j indicate the operation of taking the jth derivative, let m be a maximum order of derivative and let a set of m weight functions w_j, $j = 0, \ldots, m-1$, be available. Then L is defined as

$$L = w_0 I + w_1 D + \ldots + w_{m-1} D^{m-1} + D^m \tag{1}$$

and the value of the transformed function Ly at argument t is

$$(Ly)(t) = \sum_{j=0}^{m} w_j(t)\,(D^j y)(t).$$

The m weight functions w_j may be considered the functional parameters defining the transformation, just as the eigenfunctions v_j define a projection.

Why consider linear differential operators? First, they, also, can be used to remove components of variation. Given some mild regularity conditions (Coddington and Levinson, 1955), there are exactly m linearly independent functions u_j satisfying the homogeneous differential equation

$$Lu = 0,$$

implying that any linear combination $\Sigma_j c_j u_j$ of these functions can be annihilated by application of L. Moreover, given a target set of linearly independent u_js satisfying the same regularity conditions, it is possible to find an associated operator L that will remove them. Thus, L shares the data reduction property of a projection Q. The m-dimensional space spanned by the u_js is denoted by $\ker(L)$.

Projection and linear differential operators, although both linear, differ in terms of their images. Since Qy is in the same vector space as y, the definition of the operator identification problem as the minimization of $\|Qy\|^2$ is also in the same space. However, L is a 'roughening' transform in that Ly has m fewer derivatives than y and is usually more variable. It may be that we may want either to penalize or otherwise to manipulate y at this rough level. Put another way, it may be plausible to conjecture that the noise or unwanted variational component in y is found only at the rough level Ly. Thus, a second motivating factor for the use of L rather than Q is that the former explicitly takes account of the smoothness of the data, whereas the latter does not.

Thirdly, the identification of a linear differential operator L that removes large amounts of the signal from the data can have immediate interpretive significance in many fields. Applications in the physical sciences and engineering tend to make extensive use of differential equation models. The result $f_i = Ly_i$ is often called a *forcing or impulse function* and in physical science and engineering applications is often taken to indicate the influence of exogenous agents on the system defined by $Lu = 0$.

Finally, there is the connection with spline smoothing, for which $y_i \in R^n$. Several researchers (Ramsay and Dalzell, 1991; Ramsay *et al.*, 1995; Ansley *et al.*, 1993) have considered the general spline smoothing criterion

$$G_\lambda(h|y) = \sum_j^n \{y_j - h(t_j)\}^2 + \lambda \int (Lh)^2(t)\,\mathrm{d}t$$

where L is chosen judiciously in the light of what is known about the data being smoothed. The better-known cubic polynomial smoothing criterion employs the simple operator $L = D^2$ to penalize roughness, thus defined in terms of the behaviour of the second derivative. Spline smoothing allows a continuum of choices between projection and differential operation in the sense that in the limit $\lambda \to \infty$ observation vector y is in effect projected onto the array of values $u_j(t_i)$. But for finite λ-values the criterion penalizes a weighted sum of the projection norm used in the first term and norm of Lh used in the second term.

Intuition suggests that spline smoothing will work better if the 'default' fit to the data, when smoothing parameter λ is large, is reasonable for the data. The cubic spline criterion implies that this default fit is a straight line, and thus that cubic smoothing splines are ideal for nearly linear data.

This intuition is backed up by the asymptotic theory associated with smoothing splines. The mean-squared error (MSE) criterion

$$\mathrm{MSE} = n^{-1} \sum_j^n \{h(t_j) - g(t_j)\}^2,$$

where g is the function being estimated from the noisy sampled values y_j, has an expectation that can be expressed as

$$E[\mathrm{MSE}] = b^2(\lambda) + \sigma^2\,\mu_2(\lambda)$$

where

$$b^2(\lambda) = n^{-1} \sum_j^n \{E[h(t_j)] - g(t_j)\}^2$$

is the squared bias and the second term is the variance component, with σ^2 being the variance of sampled values. Under fairly general conditions, it can be shown (Wahba, 1990) that

$$b^2(\lambda) \leqslant \|Lg\|^2$$

whereas

$$\mu_2(\lambda) \approx O(1/n\lambda^{1/q})$$

for some $q > 1$. Thus, at least for larger n, the MSE will tend to be controlled by the squared bias, which in turn is bounded by the size of Lg. Thus, these results argue for choosing operator L to make Lg small. Moreover, if $Lg = 0$ exactly, then bias disappears with large λ and E[MSE] approaches its lower limit σ^2/n.

In the next section two techniques for PDA are presented, along with some measures of fit to the data. The possibility of regularizing or smoothing the estimated weight functions w_j is also taken up. The third section presents two examples.

2. PRINCIPAL DIFFERENTIAL ANALYSIS

Let y_i be functions differentiable up to order m and such that $D^m y \in L^2$ is square integrable over [0, 1]. Assume that $D^j y_i$, $j = 0, \ldots, m-1$, are linearly independent. By analogy with PCA, the objective is to identify a linear differential operator L such that the size of Ly is small. Once this has been achieved, the m functions $u_j \in \ker(L)$ can then be displayed by solving the homogeneous differential equation $Lu = 0$ by standard techniques.

The measure of size to be used in this paper is the L^2-norm

$$F(w|y) = \|Ly\|^2 = N^{-1} \sum_i^N \int (Ly_i)^2(t)\,dt$$

$$= N^{-1} \int (L\mathbf{y})^{\mathrm{T}}(L\mathbf{y})(t)\,dt, \tag{2}$$

where the vector-valued function $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$. Some applications for which the forcing or impulse functions are assumed to be correlated will require the generalization

$$F_\Omega(w|y) = N^{-1} \int (L\mathbf{y})^{\mathrm{T}}\, \Omega(L\mathbf{y})(t)\,dt \tag{3}$$

where order N weight matrix Ω is assumed positive definite.

2.1. *Principal Differential Analysis by Pointwise Minimization*

A pointwise estimate of the weight w_j functions is computable by standard least squares theory if we use the pointwise fitting criterion

$$F_t(w) = N^{-1} \sum_i (Ly_i)^2(t) = N^{-1} \sum_i \left\{ \sum_{j=0}^m w_j(t)(D^j y_i)(t) \right\}^2 \tag{4}$$

where, as above, $w_m(t) = 1$ for all t. Define the m-dimensional coefficient vector

$$\mathbf{w}(t) = (w_0(t), \ldots, w_{m-1}(t))^{\mathrm{T}},$$

the $N \times m$ pointwise carrier matrix

$$\mathbf{X}(t) = \{(D^j y_i)(t)\}_{i=1,\ldots,N;\, j=0,\ldots,m-1}$$

and the N-dimensional dependent variable vector

$$\mathbf{z}(t) = \{-(D^m y_i)(t)\}_{i=1,\ldots,N}.$$

Then the least squares solution minimizing F_t with respect to values $w_j(t)$ is

$$\mathbf{w}(t) = \{\mathbf{X}(t)^T\mathbf{X}(t)\}^{-1}\mathbf{X}^T(t)\mathbf{z}(t). \tag{5}$$

The existence of these pointwise values $\mathbf{w}(t)$ depends on the determinant of $\mathbf{X}(t)^T\mathbf{X}(t)$ being bounded away from 0 for all values of t, and it is wise to compute and display this determinant as a routine part of the computation. Equivalently, it is assumed that $\mathbf{X}(t)$ is of full column rank.

In fact, this pointwise solution as a function of t is also the global minimizer of criterion (2) under a slightly stronger regularity condition.

Theorem 1. If $\{\mathbf{X}^T(t)\mathbf{X}(t)\}^{-1}$ exists for all t and there exists $g \in L^2$ such that

$$\|\{\mathbf{X}^T(t)\mathbf{X}(t)\}^{-1}\| \leqslant g(t) \qquad \text{for all } t$$

then equation (5) minimizes criterion (2) over $w_j \in L^2$.

Proof. The operator

$$(\mathbf{Py})(t) = \mathbf{X}(t)\{\mathbf{X}(t)^T\mathbf{X}(t)\}^{-1}\mathbf{X}^T(t)\mathbf{y}(t)$$

is idempotent and therefore a projection onto the function space spanned by the columns of $\mathbf{X}$. Hence $\mathbf{w}$ minimizes criterion (2) within L^2. That it is itself an L^2-function follows from the Lebesgue dominated convergence theorem.

Of course, if m is not large, then the solution can be expressed in closed form. For example, for $m=1$ we have

$$w_0(t) = -\sum_i y_i(t)(Dy_i)(t) \Big/ \sum_i y_i^2(t) \tag{6}$$

and the full rank condition is merely one of requiring that for each value of t some $y_i(t)$ be non-zero. Finally, although equation (5) provides an explicit expression for $\mathbf{w}$, it should not be viewed as a recipe for computation. Better techniques for solving the least squares problem for each t involve the decomposition of $\mathbf{X}$ directly.

2.2. *Principal Differential Analysis by Basis Expansions*

The pointwise approach can pose problems in some applications. First, solving the equation $Lu = 0$ requires that the w_j be available at a fine level of detail, with the required resolution depending on their smoothness. Whether or not these functions are smooth depends in turn on the smoothness of the derivatives $D^j y_i$. Since these derivatives will often be estimated by smoothing procedures that may not always yield smooth estimates for higher order derivatives, the resolution required may be very fine indeed. But larger orders m, computing the functions w_j pointwise at a fine resolution level, can be computationally intensive, since a linear equation must be solved for every value of t for which $\mathbf{w}$ is required. This suggests the need for an approximate solution which can be quickly computed and which is reasonably regular or smooth.

Secondly, it may be desirable to circumvent the restriction in theorem 1 that the

rank of $\mathbf{X}$ be full, especially if the failure is highly localized within the interval of integration. As a rule, an isolated singularity for $\mathbf{X}^{\mathrm{T}}(t)\mathbf{X}$ corresponds to an isolated singularity in one or more weight functions w_j, and it may be desirable to bypass these by using weight functions that are sure to be sufficiently smooth. More generally, we may seek weight functions that are more smooth or regular than those resulting from the pointwise solution.

A strategy for identifying smooth weight functions w_j is to approximate them by using a fixed set of basis functions. Let ϕ_k, $k = 1, \ldots, K$, be a set of K such basis functions, and let $\boldsymbol{\phi}$ denote the K-dimensional vector function $(\phi_1, \ldots, \phi_K)^{\mathrm{T}}$. We may use standard basis families such as polynomials, Fourier series or B-spline functions with a fixed knot sequence, or we may employ a set of basis functions suggested by the application at hand. In any case, it is assumed that

$$w_j \approx \sum_k c_{jk}\phi_k \tag{7}$$

where mK-coefficients c_{jk} define the approximations and require estimation from the data. Let mK-vector $\mathbf{c}$ contain these coefficients, where index k varies inside index m.

The minimization of the criterion $F(w|y)$ with respect to $\mathbf{c}$ requires only standard linear algebra computation. Expansion of criterion (2) yields the expression

$$\hat{F}(\mathbf{c}|y) = C + \mathbf{c}^{\mathrm{T}}\mathbf{R}\mathbf{c} + 2\mathbf{c}^{\mathrm{T}}\mathbf{s} \tag{8}$$

where constant C does not depend on $\mathbf{c}$. Matrix $\mathbf{R}$ is of order mK and symmetric and consists of an $m \times m$ array of order K submatrices of the form

$$\mathbf{R}_{j_1,j_2} = \int \phi(t)\,\phi^{\mathrm{T}}(t)\,E[D^{j_1}y_i D^{j_2}y_i](t)\,\mathrm{d}t, \tag{9}$$

where

$$E[D^{j_1}yD^{j_2}y](t) = N^{-1}\sum_i^N D^{j_1}y_i\,D^{j_2}y_i(t). \tag{10}$$

Similarly, mK-dimensional vector $\mathbf{s}$ contains mK-dimensional subvectors

$$\mathbf{s}_j = \int \phi(t)\,E[D^j y D^m y](t). \tag{11}$$

Thus, $\mathbf{c}$ is the solution of the equation

$$\mathbf{s} = -\mathbf{R}\mathbf{c}. \tag{12}$$

The integrals involved in these expressions will often have to be evaluated numerically. In all the work reported in this paper, the trapezoidal rule was used over a fine mesh of equally spaced values t_s, $s = 1, \ldots, S$, $t_1 = 0$, $t_S = 1$.

2.3. *Solving $Lu = 0$*

A variety of numerical techniques for solving the homogeneous differential equation to identify the m solution functions u_j can be found in standard references on numerical methods, such as Stoer and Bulirsch (1980). To compute specific solutions,

it is necessary to specify m linear constraint conditions, and these may be of an initial value or boundary value character.

The following successive approximation procedure is simple to implement and guaranteed in principle to converge. Define order m matrix function $\mathbf{T}$ as 0 for all values of t except for

$$\mathbf{T}_{j,j+1}(t) = 1, \qquad j = 1, \ldots, m-1,$$

and

$$\mathbf{T}_{mj}(t) = -w_{j-1}(t), \qquad j = 1, \ldots, m.$$

Initialize order m matrix function $\mathbf{U}$ as $\mathbf{U}^{(0)}(t) = \mathbf{I}$, the identity matrix. On iteration $\nu \geqslant 1$ $\mathbf{U}^{(\nu)}$ is updated as follows:

$$\mathbf{U}^{(\nu)}(t) = \mathbf{I} + \int_0^t \mathbf{T}(s)\,\mathbf{U}^{(\nu-1)}(s)\,ds. \tag{13}$$

On convergence, $\mathbf{U}$ contains the solution functions u_j in the first row and their successive derivatives in subsequent rows. These solutions satisfy the initial value constraints $D^{j-1}u_j(0) = 1$. The partial integration in equation (13) can be carried out by the trapezoidal rule provided that the weight functions are available on a sufficiently fine equally spaced set of argument values. In practice, however, difficulties may be encountered with $\mathbf{U}^{(\nu)}$ becoming extremely large in size or unstable during the iterations. This can be dealt with in some cases by initializing $\mathbf{U}$ with better initial values and replacing $\mathbf{I}$ in equation (13) with a constant diagonal matrix containing more appropriate initial value constraints.

2.4. *Regularized Principal Differential Analysis*

The expansion of w_j in terms of a fixed number of basis functions can be considered a type of regularization process in that the number, choice and smoothness of the basis functions ϕ_k can be used to assure two potentially desirable features: a smooth or regular variation in w_j and the closeness of the estimated w_j to some target or hypothesized weight function ω_j.

An alternative with a proven track record in nonparametric regression is the regularization of w_j by attaching penalty terms to criterion (2). One version is

$$F(w|y) = N^{-1}\int (L\mathbf{y})^{\mathrm{T}}(L\mathbf{y})(t)\,dt + \sum_{j=0}^{m-1}\gamma_j \int \{w_j(t) - \omega_j(t)\}^2\,dt$$
$$+ \sum_{j=0}^{m-1}\lambda_j \int (D^2 w_j)^2(t)\,dt. \tag{14}$$

The scalars $\gamma_j \geqslant 0$ are smoothing parameters controlling the shrinkage in L^2-norm to target weight functions ω_j. The scalars λ_j, in contrast, control the roughness of the estimated weight functions. In fact, large values for all λ_j will shrink all the weight functions to 0, so that the differential operator will converge to $L = D^m$. A spline smoother with operator D^m is a piecewise polynomial of degree $2m-1$, so this type of regularization defines polynomial spline smoothing as a limiting case.

In spline smoothing, the roughness penalty approach is combined with a basis expansion of a very high order K, and the basis functions are usually, but not necessarily, derived from the reproducing kernel associated with a Hilbert space of functions. Wahba (1990) and Green and Silverman (1994) have discussed this topic in detail. The modification of criterion (8) to incorporate penalty terms is straightforward and will not be pursued further.

2.5. *Assessing Fit*

Since the objective of PDA is to minimize the norm $\|Ly\|$ of the forcing function associated with an estimated differential operator, and since the quality of fit can vary over the domain 1, it is appropriate to assess the fit in terms of the pointwise error sum of squares

$$\mathrm{SSE}(t) = \sum_i \|(Ly_i)(t)\|^2 = \sum_i \left\{ \sum_{j=0}^{m-1} w_j(t)(D^j y_i)(t) + (D^m y_i)(t) \right\}^2. \tag{15}$$

As in linear modelling, the logical base-line against which SSE should be compared is the error sum of squares defined by a theoretical model and its associated weight functions ω_j:

$$\mathrm{SSE}_0(t) = \sum_i \left\{ \sum_{j=0}^{m-1} \omega_j(t)(D^j y_i)(t) + (D^m y_i)(t) \right\}^2. \tag{16}$$

If there is no theoretical model at hand, we may use $\omega_j = 0$, so that the comparison is simply with the sum squares of the $D^m y_i$. From these loss functions the pointwise squared multiple correlation function

$$R^2(t) = \frac{\mathrm{SSE}_0(t) - \mathrm{SSE}(t)}{\mathrm{SSE}_0(t)} \tag{17}$$

and the pointwise F-ratio

$$F(t) = \frac{\{\mathrm{SSE}_0(t) - \mathrm{SSE}(t)\}/m}{\mathrm{SSE}_0(t)/(N-m)} \tag{18}$$

may be examined. Note, however, that the inequality $0 \leqslant R^2(t) \leqslant 1$ is only assured if the pointwise estimate of the weight functions is used.

3. ILLUSTRATIONS

3.1. *Pinch Force Data*

The data in the pinch force example consisted of 20 records of brief force impulses exerted by the thumb and forefinger in an experiment in motor physiology. These force impulses, after some preprocessing to transform time linearly to a common metric, and to remove some simple shape variation, are displayed in Fig. 1. Details concerning the experiment and the preprocessing stages can be found in Ramsay *et al.* (1995).

There are some theoretical considerations which suggest that the model

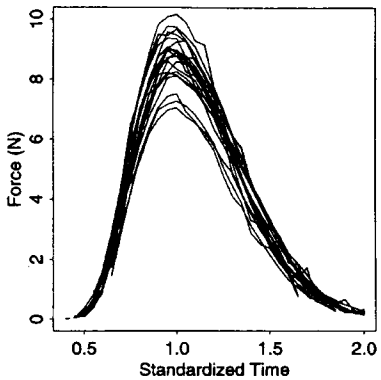

Fig. 1. 20 recordings of the force exerted by the thumb and forefinger during a brief squeeze of a force meter: the data have been preprocessed to register the functions and to remove some shape variability, and the values displayed are for the 33 values $t = 0.4, 0.45, 0.5, \ldots, 2.0$

$$y_i(t) = C_i \exp(-\log^2 t/2\sigma^2) \tag{19}$$

will offer a good account of any specific force function, and in this application the shape parameter σ is known. Functions of the form (19) are annihilated by the differential operator $L_0 = \{(t\sigma)^{-1} \log t\}I + D$. A goal of this analysis is to compare this theoretical operator with the first-order differential operator $L = w_0 I + D$ estimated from the data, or to compare the theoretical weight function $\omega_0(t) = (t\sigma)^{-1} \log t$ with its empirical counterpart w_0.

Each record was subjected to cubic spline smoothing, and the derivative function estimated by the derivative of the smoothing function. Fig. 2 displays the discrete data points, the smoothing function, its derivative and also the theoretical function (19) fitted by least squares for a single record. Note that the model comes close to the data but fails by what appears to be a consistent and smooth amount.

Both the piecewise and the global procedures were applied to the smooth functions and their derivatives. The basis used for the global procedure was

$$\phi(t) = (t^{-1} \log t,\ 1,\ t-1,\ (t-1)^2)^{\mathrm{T}},$$

chosen after some experimentation; the first basis function was suggested by the

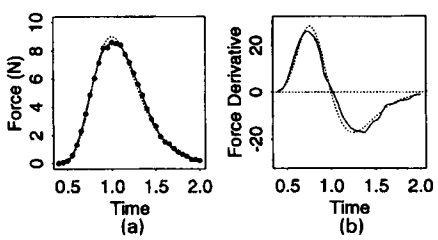

Fig. 2. (a) Data values for the first record (•), the smoothing spline (———) and the least squares fit by model (19) (·····); (b) derivative of the smoothing spline (———) and the derivative of the model (·····)

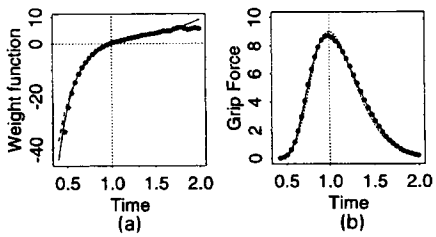

Fig. 3. (a) Globally estimated weight function for the pinch force data (———), theoretical function (- - - - - -) and piecewise estimates at selected points (•); (b) corresponding pinch force functions fit to the mean pinch force by least squares (•, selected values of the mean pinch force)

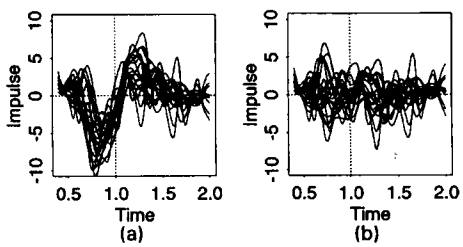

Fig. 4. (a) Forcing or impulse functions Ly_i produced by the theoretical operator ($F = 281.9$); (b) corresponding empirical operator functions ($F = 98.16$)

theoretical model, with the remaining polynomial terms extending this model as required. Fig. 3 shows the theoretical, the pointwise and the global estimates of the weight functions. The general solution to the first-order homogeneous linear equation fit to the mean force function by least squares is also displayed, along with the theoretical solution and the mean force function at selected points.

The fit of the estimated function is superior, especially near the peak force, but this is not the point of these analyses; it is the ability of the estimated operator to minimize the size of the forcing functions that matters. The values of measure F are 281.9 and 98.6 for the theoretical and empirical operators respectively, corresponding to $R^2 = 0.65$. Moreover, the forcing functions Ly_i, displayed in Fig. 4, show a systematic trend for the theoretical operator, whereas the empirical forcing functions appear not to exhibit any obvious pattern. It is appropriate to conclude that the estimated operator has produced an important improvement in fit on either side of the time of maximum force.

3.2. *Lip Motion During 'Bob'*

The next example is more exploratory in character. Ramsay *et al.* (1996) reported data on the movement of the lips during the utterance of various syllables. Their objective was to relate lip movement to possible motor control processes activating the muscles, and to assess the complexity of variation in lip behaviour over replications of utterances. For the syllable 'bob', it turned out that almost all the information was contained in the behaviour of the centre of the lower lip, which moved along a nearly linear trajectory.

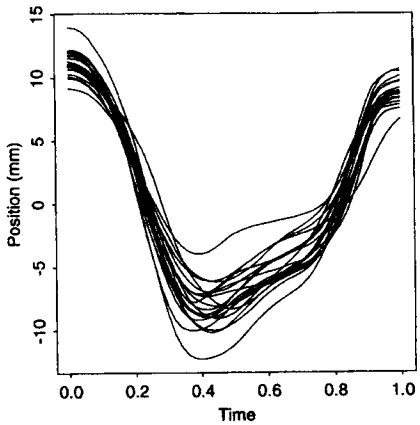

Fig. 5. 20 records of the position of the centre of the lower lip during the syllable 'bob' (the time unit is arbitrary, and the curves were preprocessed to be of fixed duration)

Fig. 5 displays 20 records of lip position relative to the mean of all values and in the direction corresponding to the principal component of variation. The time values indicated in Fig. 5 are arbitrary, but a typical utterance lasted about a third of a second. We note that the motion is typically in three phases: a first in which the lip rapidly opens, a second following exhibiting slower and nearly linear closing motion, and the third a more rapid concluding closure. The three phases are separated by two brief episodes of strong positive acceleration.

The soft tissue in the lips exhibits behaviour that reflects both its internal mechanical structure and the consequences of external control by neural activation of a number of muscle groups. It is interesting to consider how well lip motion might be described by a linear differential equation of the second order, $L = w_0 I + w_1 D + D^2 = 0$. The first coefficient, w_0, essentially reflects the position-dependent force applied to the system at position x. $w_0 > 0$ and $w_1 = 0$ correspond to a system with sinusoidal or harmonic motion, with frequency $w_0^{1/2}$ and period $w_0^{-1/2}$, and w_0 is often called the 'spring constant'. The second coefficient, w_1, indicates influences on the system that are proportional to velocity rather than position and are often internal or external frictional forces and viscosity. The *discriminant* of the system, $d = (w_1/2)^2 - w_0$, is critical in terms of its sign. When w_1 is small, so that d is negative, the system is underdamped and will tend to exhibit some visible oscillation that gradually disappears, the frequency of this oscillation being $\sqrt{(-d)}$. When d is positive because w_1 is relatively large, the system is called overdamped and will either become stable so quickly that no oscillation is observed ($w_1 > 0$) or will oscillate out of control ($w_1 < 0$). A critically damped system is one for which $d = 0$. These mechanical interpretations of the roles of coefficient functions w_0 and w_1 are strictly only appropriate if these functions are constants, but higher order effects can be ignored if they do not vary too rapidly with t, in which case $w_0(t)$, $w_1(t)$ and $d(t)$ can be viewed as describing the instantaneous state of the system. When $w_0 = w_1 = 0$ the system is in linear motion, for which $D^2 x = 0$.

The second-order equation system presumes a fixed co-ordinate system with the origin of position being known, and therefore an origin was estimated for each

record yielding the best fit. The velocity and acceleration functions, D^1y and D^2y respectively, were estimated by polynomial spline smoothing using a penalty term $\|D^4y\|$; this choice tends to produce better estimates of acceleration than the usual cubic smoothing spline penalty $\|D^2y\|$. The two weight functions were estimated by using a basis function expansion in terms of the cubic B-splines defined by the knot sequence 0, 0.05, 0.1, . . ., 1, and Fig. 6 displays their estimates. We can probably ignore the initial brief period when w_0 is negative since this is in the zone of transition from the preceding phoneme. We see that both w_0 and w_1 are small in the interval $0.5 \leqslant t \leqslant 0.8$, and that w_0 has two peaks on either side of this interval, during which the system is also unstable ($w_1 < 0$). These may correspond to the onset and termination of muscle contraction defining the central portion of small lip movement.

Two solutions u_1 and u_2 of the equation $Lu = 0$ were computed by imposing the complementary constraints $u_1(0) = 1$, $(Du_1)(0) = 0$ and $u_2(0) = 0$, $(Du_2)(0) = 1$ and solving by successive approximation. These solutions are shown in Fig. 7. The first solution has essentially the shape of the average record, whereas the second would allow for variation in slope on either side of the phase of sharp curvature at $t = 0.4$. These two functions produced excellent fits to the original data.

The discriminant function $d = (w_1/2)^2 - w_0$ is displayed in Fig. 6(b). This system is almost critically damped over the interval $0.5 \leqslant t \leqslant 0.75$, suggesting that its behaviour may be under external control. But around $t = 0.25$ and $t = 0.85$ the system is substantially underdamped and thus behaves locally like a spring. The period of the

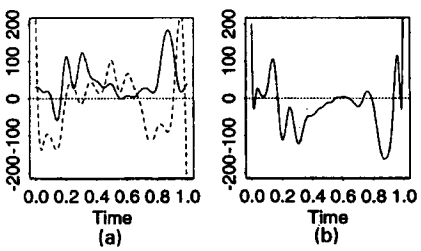

Fig. 6. (a) Weight functions w_0 (———) and w_1 (- - - - ($\times$10)) for lip motion; (b) discriminant $d = (w_1/2)^2 - w_0$

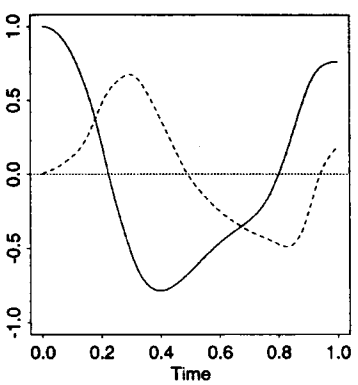

Fig. 7. Two solutions to $Lu = 0$ estimated for lip motion

spring would be around 30–40 ms, and this is in the range of values estimated in studies of the mechanical properties of flaccid soft tissue. These results suggest that the external input to lip motions tends to be concentrated in the brief period near $t = 0.6$ when the natural tendency for the lip to close is retarded to allow for the articulation of the vowel.

4. DISCUSSION AND CONCLUSIONS

PDA has been motivated in this paper as a technique for extracting the important components of variation in functional data. In this sense it shares much with PCA, and the two techniques can be viewed as being based on different linear transformation techniques. The two techniques contrast, however, in that PCA as usually understood is little concerned with smoothness properties, whereas PDA makes explicit use of smoothness in considering the behaviour of a number of levels of derivatives.

There are many ways to incorporate modelling aspects into PDA. For example, certain weight functions w_j can be fixed at predetermined values; a model for the lip motion data in which $w_1 = 0$ would allow only for purely harmonic motion. In contrast, the second-order equation

$$Lu = w_1 Du + D^2 u$$

has the general solution

$$u(t) = C_1 + C_2 \int_0^t \exp\left\{ -\int_0^s w(r)\,\mathrm{d}r \right\}$$

and is a general equation for strictly monotone twice-differentiable functions (Ramsay, 1995). This would be a natural candidate equation for modelling human growth data, as well as other applications calling for smooth monotone representations.

Considering that the computational overhead in PDA will usually be modest, and indeed rather less in many applications than that of PCA, it seems appropriate to predict an important place for this technique in both the exploratory reduction and the modelling of functional data.

ACKNOWLEDGEMENT

The author wishes to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada through grant A320.

REFERENCES

Ansley, C. F., Kohn, R. and Wong, C.-M. (1993) Nonparametric spline regression with prior information. *Biometrika*, **80**, 75–88.

Coddington, E. A. and Levinson, N. (1955) *Theory of Ordinary Differential Equations*. New York: McGraw-Hill.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. London: Chapman and Hall.

Ramsay, J. O. (1995) Estimating smooth monotone functions. To be published.

Ramsay, J. O. and Dalzell, C. J. (1991) Some tools for functional data analysis (with discussion). *J. R. Statist. Soc.* B, **53**, 539–572.

Ramsay, J. O., Munhall, K. G., Gracco, V. L. and Ostry D. J. (1996) Functional data analyses of lip motion. To be published.

Ramsay, J. O., Wang, X. and Flanagan, R. (1995) A functional data analysis of the pinch force of human fingers. *Appl. Statist.*, **44**, 17–30.

Stoer, J. and Bulirsch, R. (1980) *Introduction to Numerical Analysis*. New York: Springer.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.