Published in final edited form as:

J Stat Comput Simul. 2013 January 1; 83(10): . doi:10.1080/00949655.2012.675575.

Principal Differential Analysis with a Continuous Covariate: Low Dimensional Approximations for Functional Data

Seoweon Jin,

University of Texas at Dallas Department of Mathematical Sciences Dallas, Texas 75080-3021

Joan G. Staniswalis^{*}, and

University of Texas at El Paso Department of Mathematical Sciences El Paso, Texas 79968-0514

Indika Mallawaarachchi

Texas Tech University Health Science Center Paul L. Foster School of Medicine Department of Biostatistics & Epidemiology El Paso, Texas 79905

Abstract

Given a collection of n curves that are independent realizations of a functional variable, we are interested in finding patterns in the curve data by exploring low dimensional approximations to the curves. It is assumed that the data curves are noisy samples from the vector space $\text{span}\{f_1, \dots, f_m\}$, where $f_1, \dots, f_m$ are unknown functions on the real interval $(0, T)$ with square-integrable derivatives of all orders m or less, and $m < n$. Ramsay (1996) first proposed the method of regularized principal differential analysis as an alternative to principal component analysis for finding low dimensional approximations to curves. Principal differential analysis (PDA) is based on the following theorem: there exists an annihilating linear differential operator $\mathcal{L}$ of order m such that $\mathcal{L}f_i = 0$, $i = 1, \dots, m$ (Coddington and Levinson 1955, Theorem 6.2). Principal differential analysis specifies m , then uses the data to estimate an annihilating linear differential operator (LDO). Smooth estimates of the coefficients of the LDO are obtained by minimizing a penalized sum of the squared norm of the residuals. In this context, the residual is that part of the data curve that is not annihilated by the LDO. PDA obtains the smooth low dimensional approximation to the data curves by projecting onto the null space of the estimated annihilating LDO; PDA is thus useful for obtaining low dimensional approximations to the data curves whether or not the interpretation of the annihilating LDO is intuitive or obvious from the context of the data. This paper extends PDA to allow for the coefficients in the linear differential operator to smoothly depend upon a single continuous covariate. The estimating equations for the coefficients allowing for a continuous covariate are derived; the penalty of Eilers and Marx (1996) is used to impose smoothness. The results of a small computer simulation study investigating the bias and variance properties of the estimator are reported.

Keywords

smoothing; curve data

1 Introduction

Functional data analysis encompasses statistical methods for analysis of data obtained as a sample of a functional variable (Ramsay, Dalzell 1991; Ferraty, Vieu 2006). The functional variable maps the sample space into a Sobolev space of real-valued functions defined on the

^{*}Corresponding Author; Tel.: 00-1-915-747-6761; Fax: 00-1-915-747-5856; joan@math.utep.edu. .

open interval $(0, T)$ of the real line. The n recordings $\{x_i(t)\}_{i=1}^n$ of the independent realizations of the functional variable may be contaminated with noise and only available for t on a grid, rather than the entire interval $(0, T)$. Depending upon the number of sampling points in the interval $(0, T)$, longitudinal data analysis or time series analysis are typically used. Longitudinal data analysis (rough grid) is used to understand the mean over time, whereas, time series analysis (fine grid) focuses on understanding the spectrum when trends in the series can be estimated and removed. We are interested in finding patterns in the noisy curve data (fine grid) by exploring low dimensional approximations to the curves and their derivatives.

Principal component analysis (PCA), based on the Karhunen-Loeve expansion of the covariance function, is one approach for identifying a low-dimensional nonparametric basis for describing curve data (Rice and Silverman 1991). PCA first estimates the covariance function from the sample of curves, and then estimates the eigenfunctions in the order induced by the eigenvalues. The estimated eigenfunctions are the directions of maximal variation from the mean for the collection of curves. The low dimensional approximation of the sample is obtained by projecting each curve onto the linear space spanned by the collection of eigenfunctions that together explain most of the variation in the curve data. Estimating the derivatives of the eigenfunctions was considered by Liu and Mueller, H.G. (2006).

Principal differential analysis (PDA, Ramsay 1996) takes a very different approach, whereby it is assumed that the collection of curves can be annihilated by application of a linear differential operator (LDO). The low dimensional approximation of the curve is obtained by projecting onto the null space of the differential operator. Principal differential analysis was originally posed when a physical situation exists to justify the existence of a LDO that annihilates (or nearly annihilates) the sample of curves. One example given by Ramsay and Silverman (1997) is based on Newton's third law for mechanical systems: the release of adrenal hormones during puberty would play the analogous role of the applied force, thus allowing one to write down a second-order LDO that annihilates human growth curves. Another example given is the exponential growth of populations that satisfy a homogeneous first-order LDO. In section 2, PDA is explained and justified as useful for finding low dimensional approximations to curves and their derivatives, whether or not the existence of an annihilating LDO is intuitive or obvious for the problem at hand based on physical considerations. An extension of PDA is proposed to include an additional explanatory variable describing either the experimental units or the experimental conditions under which a recording was obtained. PCA adjusted for covariates was studied by Jiang, Aston, and Wang (2009), whereas an extension of PDA to include a single continuous covariate is new to the statistical literature and is presented in Section 3. The results of a simulation study investigating the effect of the smoothing parameters used in this extension of PDA are presented in Section 4.

2 A Linear Differential Operator

Principal differential analysis generalizes the idea of curve fitting and residuals via projections and orthogonal projections, to one of finding a linear differential operator that comes close to annihilating the noisy curve data (Ramsay and Silverman 1997).

2.1 Existence of an Annihilating Linear Differential Operator

Let $L_2(0, T)$ denote the collection of all square integrable functions on $(0, T)$ with the usual inner product and norm denoted by $\langle \cdot, \cdot \rangle_2$ and $\|\cdot\|_2$, respectively. For r a positive integer, denote by $W_2^r(0, T)$ all functions that are r -weakly differentiable (Adams 1975, page 22),

such that the k^{th} -order weak derivative $D^k f$ is square-integrable for $k = 0, \dots, r$. The Sobolev space $W_2^r(0, T)$ equipped with the inner product

$$\langle f, g \rangle_{r,2} = \int_0^T \sum_{k=0}^r (D^k f) (D^k g) dt$$

is a Hilbert space and equipped with the norm $\|f\|_{r,2} = \langle f, f \rangle_{r,2}^{1/2}$ is a Banach space (Kufner et al. 1977, page 264). Certain useful inequalities are known for Sobolev spaces and have been studied by Morosi and Pizzocchero (2006, 2010) using the norm

$$\|f\|_{MP(r,2)} = \left(\int_0^T \sum_{k=0}^r \binom{r}{k} (D^k f)^2 dt \right)^{1/2}.$$

Those inequalities carry over to the norm $\|f\|_{r,2}$, because $\|f\|_{r,2} < \|f\|_{MP(r,2)} < 2^r \|f\|_{r,2}$. For example, there exists a constant $C_{m,r,s}$ such that $\|fg\|_{m,2} \leq C_{m,r,s} \|f\|_{r,2} \|g\|_{s,2}$ for all $f \in W_2^r$ and $g \in W_2^s$, when $0 \leq m \leq r+s$ and $r+s-m > 1/2$ (Morosi, Pizzocchero 2010). An advantage of working on the Sobolev space is that unlike square-integrable functions, the Sobolev space is closed under products (set $m=r=s$; Morosi and Pizzocchero 2006).

Consider the linear differential operator $\mathcal{L}$ of order m that maps the Sobolev space $W_2^{2m}(0, T)$ into $L_2(0, T)$:

$$\mathcal{L} = w_0 I + \dots + w_{m-1} D^{m-1} + w_m D^m, \quad (2.1)$$

where $m \geq 1$, I is the identity operator, D is the weak-derivative operator and the coefficients $w_0, \dots, w_m$ are functions in $W_2^m(0, T)$. Let $M = \max\{\|w_0\|_{m,2}, \dots, \|w_m\|_{m,2}\}$. In the process of establishing that $\mathcal{L}f \in L_2(0, T)$ for $f \in W_2^{2m}(0, T)$, it is shown that the norm of $\mathcal{L}$ is bounded by $\max\{C_{0,m,2m}, \dots, C_{0,m,m+1}, C_{0,m,m}\} M(m+1)$:

$$\begin{aligned} \|\mathcal{L}f\|_2 &\leq C_{0,m,2m} \|w_0\|_{m,2} \|f\|_{2m,2} + \dots + \\ &\quad C_{0,m,m+1} \|w_{m-1}\|_{m,2} \|D^{m-1}f\|_{m+1,2} + C_{0,m,m} \|w_m\|_{m,2} \|D^m f\|_{m,2} \\ &\leq \max\{C_{0,m,2m}, \dots, C_{0,m,m+1}, C_{0,m,m}\} M(m+1) \|f\|_{2m,2}. \end{aligned}$$

Given a very large sample of curves, our goal is to estimate the coefficients of a LDO that come close to annihilating the sample of curves (this will be made more precise later). The basis functions for the null space of the estimated LDO can be used to provide a low dimensional approximation to the sample curves. However, before such a statistical analysis can be carried out, we must first justify that such a LDO exists. Referring to Coddington and Levinson (1955, Theorem 6.2), Ramsay (1996) assures us that an LDO of the form given by equation (2.1) exists under certain regularity conditions. However, the following theorem additionally establishes sufficient conditions under which the coefficients of the LDO are elements of a Sobolev space, justifying the use of nonparametric smoothing methods based on B-splines for their estimation.

Theorem—Let $f_1, \dots, f_m \in W_2^{2m}(0, T)$ be linearly independent functions, and consider $\mathcal{M} = \text{span}\{f_1, \dots, f_m\}$. Then $\mathcal{M}$ has an m^{th} order annihilating LDO of the form given by (2.1) with coefficients in $W_2^m(0, T)$.

Proof of the Theorem: The Wronskian matrix is

$$W_m(t) = \begin{bmatrix} f_1 & Df_1 & \dots & D^m f_1 \\ \vdots & \vdots & \ddots & \vdots \\ f_m & Df_m & \dots & D^m f_m \end{bmatrix}_{m \times (m+1)} \quad \text{and}$$

the conditions $\mathcal{L}f_1=0, \dots, \mathcal{L}f_m=0$ form a homogeneous system of m linear equations in the $m+1$ functions $w_0, \dots, w_m$: $W_m(t) [w_0, \dots, w_m]^t = 0$. Cramer's rule produces the solution $w_r = (-1)^{m-r} \mathbf{M}_r(t)$, for $r=0, \dots, m$, where $\mathbf{M}_r(t)$ is the determinant of the Wronskian minor obtained by deleting the $(r+1)^{\text{st}}$ column of the $m \times (m+1)$ Wronskian matrix (Kamran, Milson, and Olver 2000, page 296, example 4.2 works for functions with up to m continuous derivatives).

The solution w_r , for $r=0, \dots, m$ is a linear combination of products of elements of the Wronskian matrix $W_m(t)$: $(D^{k_1} f_{i_1}) (D^{k_2} f_{i_2}) \dots (D^{k_m} f_{i_m})$, where $0 \leq i_1, i_2, \dots, i_m \leq m$ indicate the rows and $0 \leq k_1 < k_2 < \dots < k_m \leq m$ the columns of the elements in a typical product. By repeated application of the inequalities available for the norm of a product of two functions in a Sobolev space (Morosi and Pizzocchero 2006 and 2010), there exists a constant C such that

$$\begin{aligned} \|(D^{k_1} f_{i_1}) (D^{k_2} f_{i_2}) \dots (D^{k_m} f_{i_m})\|_{m,2} &\leq C_{m,2m-k_m,2m-k_1} \|D^{k_1} f_{i_1}\|_{2m-k_1,2} \times \\ &\quad \|(D^{k_2} f_{i_2}) \dots (D^{k_m} f_{i_m})\|_{2m-k_m,2} \\ &\leq C \|D^{k_1} f_{i_1}\|_{2m-k_1,2} \|D^{k_2} f_{i_2}\|_{2m-k_2,2} \dots \|D^{k_m} f_{i_m}\|_{2m-k_m,2} \\ &\leq C \|f_{i_1}\|_{2m,2} \|f_{i_2}\|_{2m,2} \dots \|f_{i_m}\|_{2m,2}. \end{aligned}$$

The proof is now complete, because $f_1, \dots, f_m \in W_2^{2m}$. B-splines of order $p = m+1$ provide a convenient basis for the approximation of elements from the Sobolev space W_2^m (de Boor 1978, Eubank 1988). So we have justified the approximation of the coefficients $w_0, \dots, w_m$ by B-spline basis functions.

2.2 B-Splines

A spline function of order p on an interval $[0, T]$ with real-valued knot sequence $0 < T_1 < \dots < T_k < T$ is given by

$$S(t; \theta, \delta) = \sum_{i=0}^{p-1} \theta_i t^i + \sum_{i=1}^k \delta_i (t - T_i)_+^{p-1},$$

with coefficients $\theta = (\theta_0, \dots, \theta_{p-1}) \in \mathcal{R}^p$ and $\delta = (\delta_1, \dots, \delta_k) \in \mathcal{R}^k$. Here $(\cdot)_+ = \max\{\cdot, 0\}$. This definition of a spline function $S(t; \cdot, \cdot)$, $t \in [0, T]$, uses the truncated power basis. The collection of all order p spline functions on the interval $[0, T]$ with knot sequence $\{T_i\}_{i=1}^k$ is a vector space of dimension $p+k$ over the real numbers. B-splines provide a convenient basis for the representation of elements from this vector space. Define $2p$ additional knots by $T_{-(p-1)} = \dots = T_{-1} = T_0 = 0$, and $T_{k+1} = \dots = T_{k+p} = T$. The i^{th} normalized B-spline of order p is given by the recursion:

$$N_{i,p}(t) = \frac{(t - T_i)}{(T_{i+p-1} - T_i)} N_{i,p-1}(t) + \frac{(T_{i+p} - t)}{(T_{i+p} - T_{i+1})} N_{i+1,p-1}(t),$$

for $i = -(p-1), \dots, k$. The recursion starts with $N_{j,1}(t)$ equal to the indicator function of the interval $[T_j, T_{j+1})$ for $j = 0, \dots, k$. A spline of order p with $p+1$ coincident knots is by definition the zero function. The compact local support of the B-splines: $N_{i,p}(t) = 0$, if $t \notin [T_i, T_{i+p}]$, motivates the definition of P-splines (Eilers and Marx 1996) explained in the next section. Let the B-spline basis of dimension $I = p + k$ on $(0, T)$ be denoted by $\{N_1, N_2, \dots, N_I\}$

2.3 Estimating the Coefficients in the Linear Differential Operator

Ramsay (1996) studies the linear differential operator

$$\langle no \rangle (2.2) \langle /no \rangle \mathcal{L} = w_0 I + \dots + w_{m-1} D^{m-1} + D^m$$

obtained by setting $w_m(t) = 1$ in equation (2.1). In this case, the solution to the coefficients of the LDO that annihilates $\text{span}\{f_1, \dots, f_m\}$ is $w_r = (-1)^{m-r} \mathbf{M}_r(t) / \mathbf{M}_{m+1}(t)$, for $r = 0, \dots, m-1$, where $\mathbf{M}_{m+1}(t)$ is the determinant of the Wronskian minor obtained by deleting the $(m+1)^{\text{th}}$ column of the Wronskian matrix. It is no longer clear that the coefficients of the annihilating LDO (2.2) are in the Sobolev space W_2^m , because the Sobolev space is not closed under division. Nevertheless, we will proceed with estimation of $w_0, \dots, w_{m-1}$ in (2.2) as in Ramsay and Silverman (1997; 2002) using B-spline basis approximations.

Approximate the coefficients w_j of the linear operator $\mathcal{L}$ by $\sum_{i=1}^I \beta_{ji} N_i$, and stack all the B-spline coefficients in a vector: $\beta^T = (\beta_0^T, \beta_1^T, \dots, \beta_{m-1}^T)$, where $\beta_j^T = (\beta_{j1}, \dots, \beta_{jI})$ for $j=0, \dots, m-1$. The fitting criterion for estimating the vector of coefficients is described next.

The operator $\mathcal{L}$ should annihilate the sample of curves $\{x_i\}_{i=1}^n$ as nearly as possible. As the fitting criterion, Ramsay and Silverman (2002) use the sum of squared norms

$$\langle no \rangle (2.3) \langle /no \rangle \sum_{i=1}^n \int_0^T (\mathcal{L} x_i(t))^2 dt$$

and suggest a regularized version obtained by adding a penalty term. We prefer to use the penalty for smoothness based on Eilers and Marx (1996): $\sum_{j=0}^{m-1} \lambda_j \|D^2 \beta_j\|^2$. Here it is understood from the context that D acting on a vector is the first-order differencing operator. Thus the criterion is to find $\hat{\beta}$ that minimizes

$$Q(\beta) = \sum_{i=1}^n \int_0^T (\mathcal{L} x_i(t))^2 dt + \sum_{j=0}^{m-1} \lambda_j \|D^2 \beta_j\|^2$$

with respect to β . The penalty term is using second-order finite differences of the coefficients of adjacent B-spline basis functions that have overlapping support. The difference penalty has the effect of imposing smoothness by requiring that the heights

(coefficients) of neighboring B-splines “hold hands” to avoid erratic fluctuations in the fitted curve (Eilers and Marx 1996). Then the estimators of the coefficient functions in the LDO

are taken to be $\hat{w}_j \approx \sum_{i=1}^I \hat{\beta}_{ji} N_i$.

2.4 Low Dimensional Approximations of $X(t)$ Using PDA

The null space is the set of all elements in the Sobolev space mapped to zero by the LDO. Generally, there are m linearly independent functions, say $U_1(t), \dots, U_m(t)$, that form a basis of the null space corresponding to a linear differential operator $\mathcal{L}$ of order m as given by equation (2.1). The Runge-Kutta (Press et al 1992) method is used to compute the basis functions. The basis functions are uniquely determined because of the nature of the initial conditions imposed on the solutions obtained by numerical iteration. Let the $m \times m$ matrix $U(t)$ contain the solution functions $U_1(t), \dots, U_m(t)$ in the first row, and their successive derivatives in subsequent rows. The m -linearly independent initial conditions for the iteration are set to the matrix $U(t)|_{t=0}$ equal to the identity matrix.

Given $x \in W_2^{2m}(0, T)$, using the inner product $\langle \cdot, \cdot \rangle_{2m,2}$ of the Sobolev space, the projection $Px = \sum_{j=1}^m a_j U_j(t)$ onto the null space of $\mathcal{L}$ provides a low dimensional approximation. Note that Ramsay and Silverman (1997) instead obtain the low dimensional approximation by projecting $x(t)$ onto $\text{span}\{U_1(t), \dots, U_m(t)\}$ using the inner product $\langle \cdot, \cdot \rangle_2$ for $L_2(0, T)$, but that choice does not consider the fidelity to derivatives of the curve.

3 A Linear Differential Operator with a Single Continuous Covariate

In many applications the functional data are collected from heterogeneous experimental units that have a covariate or explanatory variable measured. So we want to expand the idea of the previous section to include a continuous covariate. Suppose that we have n functional observations $x_1, \dots, x_n$ with corresponding covariate values $v_1, \dots, v_n$ respectively. Let $\mathcal{L}_v$ be the linear differential operator defined as

$$\mathcal{L}_v = w_0(\cdot, v)I + w_1(\cdot, v)D + \dots + w_{m-1}(\cdot, v)D^{m-1} + D^m, \quad (3.4)$$

where $w_j(\cdot, v)$ is now a function of both t and v for $j = 0, \dots, m-1$. As before, the fitting criterion remains

$$\sum_{i=1}^n \int_0^T (\mathcal{L}_v x_i(t))^2 dt,$$

where $\mathcal{L}_i = \mathcal{L}_{v_i}$ ($i=1, \dots, n$). We assume that $w_j(t, v)$ is a product of a function in t and a function in v that may be well approximated as follows:

$$w_j(t, v) \approx \left(\sum_{i'=1}^I \alpha_{ji'} N_{i'}(t) \right) \left(\sum_{l=1}^L \beta_{jl} B_l(v) \right), \quad (3.5)$$

where $\{B_1(v), \dots, B_L(v)\}$ is a B-spline basis of order p and dimension L for a given knot sequence defined on the support of the covariate v . Let $\beta_{jl} = \beta_{ji} \beta_{jk}$ then β_{ji} are identifiable, not β_{jk} and β_{jl} . This assumption (3.5) is reasonable for one covariate v whenever $w(t, v)$

(suppressing the j) is approximated well by $\sum_{i'=1}^I \alpha_{i'}(v) N_{i'}(t)$ and $\beta_{ji}(v)$ is approximated

well by $\sum_{l=1}^L \theta_{l'} B_l(v)$, for example, in modeling the effect of age (v) on the cortical auditory evoked potentials in response to speech sound stimuli (Tremblay and Ross 2007).

Let $\mathbf{N}_t = (N_1(t), \dots, N_L(t))^T$, $\mathbf{B}_v = (B_1(v), \dots, B_L(v))^T$, $\mathbf{N}_t \otimes \mathbf{B}_v = \begin{pmatrix} N_1(t) \mathbf{B}_v \\ \vdots \\ N_L(t) \mathbf{B}_v \end{pmatrix}$. Set $\beta_j = (\beta_{j1}, \dots, \beta_{jL})^T$, $\beta_j = (\beta_{j1}, \dots, \beta_{jL})^T$, $(j=0, \dots, m-1)$, $\alpha = (\alpha_0^T, \dots, \alpha_{m-1}^T)^T$, $\beta = (\beta_0^T, \dots, \beta_{m-1}^T)^T$, and $\alpha \otimes \beta = ((\alpha_0 \otimes \beta_0)^T, \dots, (\alpha_{m-1} \otimes \beta_{m-1})^T)^T$. Smoothness is imposed on $w_j (j=0, \dots, m-1)$ by adding two penalty terms; one term differences the β_j and the other one differences the α_j . The usual penalty term for β_j is $\beta_j^T D^2 \beta_j$ and the usual penalty term for α_j is $\alpha_j^T D^2 \alpha_j$ so the following penalized fitting criterion at first seems like the obvious choice:

$$\sum_{i=1}^n \int_0^T (\mathcal{L}_i x_i(t))^2 dt + \sum_{j=0}^{m-1} \gamma_j \|D^2 \beta_j\|^2 + \sum_{j=0}^{m-1} \lambda_j \|D^2 \alpha_j\|^2,$$

but it is not convenient to work with; it can be shown that $\alpha_j \otimes \beta_j$ and $\beta_j \otimes \alpha_j$ appear in the different terms as $\alpha_j \otimes \beta_j$, alone, and $\beta_j \otimes \alpha_j$ alone.

Instead, we choose to work with the penalty $\sum_{j=0}^{m-1} \gamma_j \|D^2 (\alpha_j \otimes \beta_j)\|^2$ that is in effect differencing the β_j and $\sum_{j=0}^{m-1} \lambda_j \|D^2 (\beta_j \otimes \alpha_j)\|^2$ that differences the α_j . It will be shown that the following penalized criterion can be minimized with respect to $\alpha \otimes \beta$:

$$(3.6) \quad Q(\alpha, \beta) = \sum_{i=1}^n \int_0^T (\mathcal{L}_i x_i(t))^2 dt + \sum_{j=0}^{m-1} \gamma_j \|D^2 (\alpha_j \otimes \beta_j)\|^2 + \sum_{j=0}^{m-1} \lambda_j \|D^2 (\beta_j \otimes \alpha_j)\|^2.$$

By substitution of (3.4) and (3.5) into (3.6), and expressing the result using matrix notation:

$$\begin{aligned}
\mathcal{L}_v x(t) &= \begin{pmatrix} \alpha_0^T, \dots, \alpha_{m-1}^T \end{pmatrix} \begin{pmatrix} \left(\sum_{l=1}^L \beta_{0l} N_1(t) B_l(v) \right) \cdot x(t) \\ \vdots \\ \left(\sum_{l=1}^L \beta_{0l} N_1(t) B_l(v) \right) \cdot x(t) \\ \vdots \\ \left(\sum_{l=1}^L \beta_{m-1l} N_1(t) B_l(v) \right) \cdot D^{m-1} x(t) \\ \vdots \\ \left(\sum_{l=1}^L \beta_{m-1l} N_1(t) B_l(v) \right) \cdot D^{m-1} x(t) \end{pmatrix} + D^m x(t) \\
&= \left((\alpha_0 \otimes \beta_0)^T, \dots, (\alpha_{m-1} \otimes \beta_{m-1})^T \right) \begin{pmatrix} (\mathbf{N}_t \otimes \mathbf{B}_v) \cdot x(t) \\ \vdots \\ (\mathbf{N}_t \otimes \mathbf{B}_v) \cdot D^{m-1} x(t) \end{pmatrix} + D^m x(t) \\
&= \mathcal{L}_{v,m-1} x(t) + D^m x(t).
\end{aligned}$$

Setting $\mathcal{L}_{i,m-1} = \mathcal{L}_{v_i,m-1}$ for $i = 1, \dots, n$,

$$\begin{aligned}
Q(\alpha, \beta) &= \sum_{i=1}^n \int_0^T (\mathcal{L}_{i,m-1} x_i(t)) (\mathcal{L}_{i,m-1} x_i(t))^T dt + 2 \sum_{i=1}^n \int_0^T (\mathcal{L}_{i,m-1} x_i(t)) D^m x_i(t) dt \\
&\quad + \sum_{i=1}^n \int_0^T (D^m x_i(t))^2 dt + (\alpha \times \beta)^T \mathbf{H}_1^T \mathbf{H}_1 (\alpha \times \beta) + (\alpha \times \beta)^T \mathbf{H}_2^T \mathbf{H}_2 (\alpha \times \beta),
\end{aligned}$$

and denoting by C the terms that do not involve α and β :

$$\begin{aligned}
(3.7) \quad Q(\alpha, \beta) &= C + (\alpha \times \beta)^T \mathbf{D} (\alpha \times \beta) + 2(\alpha \times \beta)^T \mathbf{d} \\
&\quad + (\alpha \times \beta)^T \mathbf{H}_1^T \mathbf{H}_1 (\alpha \times \beta) + (\alpha \times \beta)^T \mathbf{H}_2^T \mathbf{H}_2 (\alpha \times \beta).
\end{aligned}$$

Differentiate the above equation (3.7) with respect to $\alpha \times \beta$, then the estimates of $\alpha \times \beta$ satisfy the equation

$$(3.8) \quad (\mathbf{D} + \mathbf{H}_1^T \mathbf{H}_1 + \mathbf{H}_2^T \mathbf{H}_2) (\alpha \times \beta) = -\mathbf{d}.$$

The symmetric matrix $\mathbf{D}$ in (3.7) is of order mL , and consists of an $m \times m$ array of $IL \times IL$ submatrices $\mathbf{D}_{jl}$ of the form

$$\mathbf{D}_{jl} = \int_0^T \sum_{i=1}^n D^{j-1} x_i(t) D^{l-1} x_i(t) (\mathbf{N}_t \otimes \mathbf{B}_{v_i}) (\mathbf{N}_t \otimes \mathbf{B}_{v_i})^T dt$$

for $j, l = 1, \dots, m$. Similarly, the vector $\mathbf{d}$ in (3.7) consists of m subvectors $\mathbf{d}_j$ of length IL defined as

$$\mathbf{d}_j = \int_0^T \sum_{i=1}^n D^{j-1} x_i(t) D^m x_i(t) (\mathbf{N}_t \otimes \mathbf{B}_{v_i}) dt$$

for $j = 1, \dots, m$. The matrix $\mathbf{H}_1$ in (3.7) is a $m(\text{IL} - 2) \times m\text{IL}$ matrix, and consists of an $m \times m$ array of $(\text{IL} - 2) \times \text{IL}$ submatrices $\mathbf{H}_{jl}^{(1)}$ of the form

$$\mathbf{H}_{jl}^{(1)} = \begin{cases} \sqrt{\gamma_{j-1}} \mathbf{H}_0^{(1)} & \text{if } j=l \\ \mathbf{O} & \text{otherwise} \end{cases}$$

for $j, l = 1, \dots, m$, and $\mathbf{H}_0^{(1)} = (H_{h,k})$ is a $(\text{IL} - 2) \times \text{IL}$ matrix of the form

$$H_{h,k} = \begin{cases} 1 & \text{if } k=h, h+2 \\ -2 & \text{if } k=h+1 \\ 0 & \text{otherwise} \end{cases}$$

for $h = 1, 2, \dots, \text{IL} - 2$; $k = 1, 2, \dots, \text{IL}$. The matrix $\mathbf{H}_2$ in (3.7) is a $m(\text{IL} - 2) \times m\text{IL}$ matrix, and consists of an $m \times m$ array of $(\text{IL} - 2) \times \text{IL}$ submatrices $\mathbf{H}_{jl}^{(2)}$ of the form

$$\mathbf{H}_{jl}^{(2)} = \begin{cases} \sqrt{\lambda_{j-1}} \mathbf{H}_0^{(2)} & \text{if } j=l \\ \mathbf{O} & \text{otherwise} \end{cases}$$

for $j, l = 1, \dots, m$, and $\mathbf{H}_0^{(2)} = \mathbf{H}_0^{(1)} \mathbf{P}$ is a $(\text{IL} - 2) \times \text{IL}$ matrix, where $\mathbf{P}$ is an orthogonal matrix of order IL satisfying $\begin{pmatrix} j & j \end{pmatrix} = \mathbf{P} \begin{pmatrix} j & j \end{pmatrix}$.

Once the coefficient functions $w_0(t, v), \dots, w_{m-1}(t, v)$ are estimated, estimates of the linear differential operators $\mathcal{L}_v$ are obtained. By fixing v , the problem of computing the basis functions and low dimensional approximation for a curve $x(t)$ with covariate value v reduces to the computations in the previous section for PDA without covariates.

4 Simulations

In this section, the effect of the smoothing parameters in the penalized fitting criterion is investigated for $m = 2$ when the smoothing parameters are restricted to $\lambda_0 = \lambda_1 = \lambda$ and $\gamma_0 = \gamma_1 = \gamma$. Fifty simulated data sets were generated, the smoothing parameter values for λ and γ were varied over a grid and the same fifty data sets were analyzed with each pair of smoothing parameter values. The pointwise mean and variance computed from the fifty estimates is used to report bias and variance for estimators of the basis functions: U_1 and U_2 . Recall that these estimated basis functions are computed only after first obtaining $\hat{w}_0(t, v)$ and $\hat{w}_1(t, v)$ that minimize the penalized sum of the squared-norm of residuals obtained by applying the LDO to the data curves. The effect of the smoothing parameters (λ and γ) on $\hat{U}_1(t, v)$ and $\hat{U}_2(t, v)$ enters in through the estimators $\hat{w}_0(t, v)$ and $\hat{w}_1(t, v)$, so that the bias and variance properties of $\hat{U}_1(t, v)$ and $\hat{U}_2(t, v)$ should be directly inherited from the corresponding properties of $\hat{w}_0(t, v)$ and $\hat{w}_1(t, v)$.

The coefficients of a second order differential operator of the form studied by Ramsay (1996) and given by equation (2.2) were fixed at

$$w_0(t, v) = 10v^2, w_1(t, v) = 2v \text{ for } t, v \in [0, 2].$$

This second order differential equation annihilates the curves $e^{-vt}(a_1 \cos(3vt) + a_2 \sin(3vt))$. The basis functions that satisfy the initial conditions $U_1(0, v) = 1, U_2(0, v) = 0, U_1'(0, v) = 0, U_2'(0, v) = 0$ are

$$\begin{aligned} U_1(t, v) &= e^{-vt} \cos(3vt) + \frac{1}{3} e^{-vt} \sin(3vt) \\ U_2(t, v) &= \frac{1}{3v} e^{-vt} \sin(3vt). \end{aligned}$$

The functional data were randomly generated according to

$$x_i(t, v_i) = e^{-v_i t} (a_{1,i} \cos(3v_i t) + a_{2,i} \sin(3v_i t)) + \sigma \epsilon_i(t),$$

where $\{\epsilon_i\}_{i=1}^n$ is an iid sample from a standard normal, $\{a_{1,i}, a_{2,i}\}_{i=1}^n$ are iid according to a bivariate normal distribution with mean $\begin{pmatrix} 2 \\ 2 \end{pmatrix}$ and covariance $\begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$, and independently sampled from $\{\epsilon_i(t)\}_{i=1}^n$. The simulations were first run by setting $\lambda = 0$, then the simulations were run again with the same seed and $\lambda = .15$ to understand the effect of sampling noisy curve data. Here the covariate values were set to $v_i = \frac{2i}{50}, i = 1, \dots, 50$, and the curves were sampled on the regular grid $t_j = \frac{j-1}{100}$, for $j = 1, \dots, 200$.

The sampled data curves were preprocessed with light smoothing using cubic P-splines. When $\lambda = 0$, the smoothing parameter in the preprocessing step was set to 10^{-5} using a B-spline basis of arbitrarily chosen dimension 40 to interpolate each data curve; when $\lambda = 0.15$, the smoothing parameter was increased to 10. The coefficients $\hat{w}_0(t, v), \hat{w}_1(t, v)$ for each v were computed according to Section 3 in a R-software implementation using cubic B-spline bases of dimension $I = L = 20$ for regularly spaced knots in t and v , respectively. The analysis was carried out for the quartiles of the covariate: $v = 0.5, 1, 1.5$. Then the numerical algorithm available at Dr. Ramsay's website "<http://ego.psych.mcgill.ca/misc/fda>" was used to obtain the estimates $\{\hat{U}_1(t, v), \hat{U}_2(t, v)\}$ for a null space basis of $\widehat{\mathcal{L}}_v$.

Tables 1a and 1b have the tabulated values of average variance and average squared-bias of $\hat{U}_1(t, v)$ and $\hat{U}_2(t, v)$, respectively, from which a subset was selected for plotting. The averages in Tables 1(a) and 1(b) refer to averages over both $t \in \{0, .01, \dots, 1.99\}$ and $v \in \{.5, 1, 1.5\}$ of the pointwise squared-bias and variance as estimated from the simulations. Average mean-squared error is dominated by the bias-squared component, which is of order 10^{-3} , because variance is only of order 10^{-4} .

Let's first examine the relationship displayed in Figure 1 ($\lambda = 0$) between average variance (column 1) and the smoothing parameters λ and μ . The average variance is decreasing in λ , but increasing in μ . Note that λ is smoothing over correlated observations in time within a data curve, while μ is smoothing across covariate values identifying independent data curves. Hart and Wehrly (1986) discussed the properties of scatterplot smoothers applied to correlated data, pointing out that asymptotic mean-squared error for smoothing over

correlated data can be quite different from that obtained for uncorrelated data; this difference is driven by the variance component of the mean-squared error and depends upon the smoothness of the covariance function $\text{Cov}(x(s), x(t))$ along a neighborhood of the diagonal $s = t$. However, the results of Hart and Wehrly (1986) are not directly applicable here, because they did not have any covariates attached to the independent trajectories, allowing them to average over all the trajectories, whereas we use $\hat{U}$ to smooth across data curves with different values of v .

Next, the patterns observed in the simulations for average squared-bias with $\gamma = 0$ (Figure 1, second column) are described. The behavior of ordinary scatterplot smoothers of polynomial functions, such as $w_0(t, v) = 10v^2$ and $w_1(t, v) = 2v$ that are constant in t and quadratic or linear in v , are described by Hart and Wehrly (1992). Namely, for scatterplot smoothing of constant functions in t , we expect to be able to use a large smoothing parameter γ without suffering the effects of a large bias from either smoothing or boundary problems; for scatterplot smoothing of a linear function in v , we expect that increasing the smoothing parameter γ will increase bias from the poor boundary behavior of the estimator; for scatterplot smoothing of a quadratic function in v , we expect that increasing the smoothing parameter γ will increase bias from both the poor boundary behavior and bias from smoothing. The described relationships between the squared-bias of the estimators of the coefficients of the LDO, $w_0(t, v) = 10v^2$ and $w_1(t, v) = 2v$, and the smoothing parameters are inherited by $\hat{U}_1$ and $\hat{U}_2$ with an exception occurring when oversmoothing in t using the largest value of $\gamma = 2.5 \times 10^7$.

Figure 2 summarizes the results of the simulation when the smooth curves are contaminated with noise ($\gamma = .15$). As before, the variance (first column) is decreasing in γ , but here the relationship between variance and γ switches from decreasing in γ (when undersmoothing in $t = .25$) to increasing in γ (when oversmoothing in $t = 2.5 \times 10^7$).

In order to confirm the behavior of average variance as a function of γ , Figures 3a, 3b ($\gamma = 0$) and 4a, 4b ($\gamma = .15$) display the 50 realizations of the estimators for $\hat{U}_1$ (row 1) and $\hat{U}_2$ (row 2) with $v = 1$. Figures 3a and 4a display the estimates for undersmoothing in t ($\gamma = .25$), and Figures 3b and 4b for oversmoothing ($\gamma = 2.5 \times 10^7$). When undersmoothing in t ($\gamma = 2.5$), the pointwise variance of the estimators is decreasing in γ , but then changes to increasing in γ ; see Figures 3a and 4a. When γ is oversmoothing ($\gamma = 2.5 \times 10^7$), pointwise variance is increasing in γ ; see Figures 3b and 4b. In summary, the average variance is increasing with smoothing over the covariate v , except when undersmoothing in t .

Finally, we study the pointwise bias of $\hat{U}_1$ and $\hat{U}_2$ as a function of γ with $\gamma = .25$ (Figure 5a with $\gamma = 0$; Figure 6a for $\gamma = .15$), and as a function of γ with $\gamma = 25$ (Figure 5b with $\gamma = 0$; Figure 6b for $\gamma = .15$). In these figures, the rows refer to the basis functions $\hat{U}_1$ and $\hat{U}_2$ and the columns are the value of the covariate, which were set to $v = .5, 1.0, 1.5$. Some patterns that emerge are that the bias is increasing in γ , and is largest for the boundary values of $v = .5, 1.5$ as compared with $v = 1.0$. For the moment focusing on $v = 1$ to ignore boundary effects, increasing γ (smoothing in t) up to a point can serve to decrease bias, after which bias increases (Figure 6b, column 2). The smoothing parameter γ has the smaller effect on bias as compared with γ (compare Figures 5a and 5b, as well as Figures 6a and 6b). This is consistent with the earlier discussion of the bias properties of scatterplot smoothers of constant and polynomial functions, upon noting that the coefficients w_0 and w_1 were constant functions in t , but linear and quadratic functions in v .

5 Conclusion

An extension of principal differential analysis that allows for inclusion of a single continuous covariate has been provided. A simulation study was used to study the bias and variance properties of the estimators of the basis functions for the null space of the linear differential operator. Increasing smoothing on the variable v (pooling information across subjects with correlated observations in t) could result in a larger variance; others have reported this property when smoothing over correlated data. Analysis is needed to obtain asymptotic expressions for the bias and variance of the basis function estimators to understand this behavior.

Any collection of m linearly independent functions in W_2^{2m} has an annihilating LDO of the form given by equation (2.1) with coefficients in W_2^m ; this justifies the approximation of the coefficient functions by a B-spline basis. The open-source implementation of Ramsay (1996) assumes that $w_m(t) = 1$ by dividing the coefficients through by $w_m(t)$, but the Sobolev space of functions is not closed under division. Thus, theoretical justification for the B-spline approximation for the coefficient functions is lost. Further work is needed to obtain an implementation of PDA for equation (2.1) and to determine the practical effects of using Ramsay's existing implementation (used here) of (2.2).

The findings here are encouraging and suggest further development is needed in extending PDA to include information from several covariates. The approach considered here was computationally intensive and involved inverting large matrices. This may not be feasible with several covariates; the simplification that the coefficient functions are products of functions in t alone and in v alone may be too restrictive.

Acknowledgments

We thank Professor Peter J. Olver, University of Minnesota, Minneapolis, for his guidance in the proof of the theorem; the authors are responsible for any errors. The authors's effort on this project was partially supported by NIH-SCORE 2S06 GM08012 and NIH-1S3GM094073-1.

References

- [1]. Adams, RA. Sobolev Spaces. Academic Press; Massachusetts: 1975.
- [2]. Coddington, EA.; Levinson, N. Theory of Ordinary Differential Equations. McGraw-Hill; NY: 1955.
- [3]. de Boor, C. A Practical Guide to Splines. Springer-Verlag; New York: 1978.
- [4]. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. Statistical Science. 1996; 11(2):89–121.
- [5]. Eubank, RL. Spline Smoothing and Nonparametric Regression. Marcel Dekker; New York: 1988.
- [6]. Ferraty, F.; Vieu, P. Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics; 2006.
- [7]. Hart JD, Wehrly TE. Kernel regression estimation using repeated measurements data. Journal of the American Statistical Association. 1986; 81:1080–1088.
- [8]. Hart JD, Wehrly TE. Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models. Journal of the American Statistical Association. 1992; 87:1018–1024.
- [9]. Jiang CR, Aston JAD, Wang JL. Smoothing dynamic positron emission tomography time courses using functional principal components. NeuroImage. 2009; 47:184–193. [PubMed: 19344774]
- [10]. Kamran N, Milson R, Olver PJ. Invariant modules and the reduction of nonlinear partial differential equations to dynamical systems. Advances in Mathematics. 2000; 156:286–319.

- [11]. Kufner, A.; Oldrich, John, O.; Fucik, S. Function Spaces. The Netherlands, Noordho International Publishing; Leyden: 1977.
- [12]. Liu B, Mueller HG. Estimating derivatives for samples of sparsely observed functions, with application to on-line auction dynamics. *J. of the American Statistical Assoc.* 2006; 104:707–717.
- [13]. Morosi C, Pizzocchero L. On the constants for multiplication in Sobolev spaces. *Advances in Applied Mathematics.* 2006; 36:319–363.
- [14]. Morosi C, Pizzocchero L. New results on multiplication in Sobolev spaces. *Advances in Applied Mathematics.* 2010; 44:393–432.
- [15]. Press, WH.; Teukolsky, SA.; Vetterling, WT.; Flannery, BP. Numerical recipes in C. Second edition. Cambridge University Press; 1992.
- [16]. Ramsay JO, Dalzell CJ. Some tools for functional data analysis. *Journal of the Royal Statistical Society.* 1991; 53(3):539–572. Series B
- [17]. Ramsay JO. Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society.* 1996; 58:495–508. Series B
- [18]. Ramsay, JO.; Silverman, BW. Functional Data Analysis. Springer; New York: 1997.
- [19]. Ramsay, JO.; Silverman, BW. Applied Functional Data Analysis. Springer; New York: 2002.
- [20]. Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society.* 1991; 53:233–243. Series B
- [21]. Tremblay KR, Ross B. Effects of age and age-related hearing loss on the brain. *Journal of Communication Disorders.* 2007; 40:305–312. [PubMed: 17433356]

Figure 1.

Simulated Variance and Squared-Bias averaged over t and the covariate v for various values of λ and γ , when the data curves are not contaminated with noise ($\sigma = 0$).

Figure 2.

Simulated Variance and Squared-Bias averaged over t and the covariate v for various values of λ and γ , when the data curves are contaminated with noise ($\sigma = .15$).

Figure 3(a).

Simulated estimates of the basis functions for various values of γ with $\alpha = .25$, when the data curves are not contaminated with noise ($\epsilon = 0$, $v = 1$).

Figure 3(b).

Simulated estimates of the basis functions for various values of γ with $\sigma = 2.5 \times 10^7$, when the data curves are not contaminated with noise ($\epsilon = 0$, $v = 1$).

Figure 4(a).

Simulated estimates of the basis functions for various values of γ with $\alpha = .25$, when the data curves are contaminated with noise ($\sigma = 0.15$, $\nu = 1$).

Figure 4(b).

Simulated estimates of the basis functions for various values of γ with $\gamma = 2.5 \times 10^7$, when the data curves are contaminated with noise ($\sigma = 0.15$, $\nu = 1$).

Figure 5(a).

Simulated Pointwise Bias for various values of v with $\gamma = .25$, when the data curves are not contaminated with noise ($\sigma = 0$).

Figure 5(b).

Simulated Pointwise Bias for various values of v with $\lambda = 25$, when the data curves are not contaminated with noise ($\sigma = 0$).

Figure 6(a). Simulated Pointwise Bias for various values of γ with $\nu = .25$, when the data curves are contaminated with noise ($\sigma = 0.15$).

Figure 6(b). Simulated Pointwise Bias for various values of λ with $n = 25$, when the data curves are contaminated with noise ($\sigma = 0.15$).

$10^4 \times$ Average Variance and $10^3 \times$ Average Squared-Bias of $\hat{U}_1$ for Values of ϵ and δ on a Grid with $\epsilon = 0$. Values listed in parenthesis are for $\delta = 0.15$, but are not reported for values of $\delta < .25$, which undersmoothed in t when the data curves were contaminated with noise.

Table 1a

$10^4 \times$ Variance	0.0025	0.0125	0.025	0.25	25	2500	2.5×10^4	2.5×10^5	2.5×10^7
.0025	49	14	9	5 (210)	3 (66)	1 (61)	0.8 (45)	0.5 (32)	3 (21)
0.25	24	8	6	3 (72)	2 (29)	1 (38)	1 (40)	4 (33)	7 (23)
2.5	33	12	9	9 (43)	8 (18)	6 (28)	7 (31)	10 (32)	16 (34)
25	28	14	13	17 (33)	18 (21)	15 (30)	16 (33)	16 (33)	21 (41)
250	42	33	35	49 (37)	40 (40)	26 (44)	26 (48)	26 (49)	31 (54)
2500	58	36	37	53 (33)	45 (44)	40 (62)	42 (68)	42 (69)	42 (70)

$10^3 \times$ Squared-Bias	0.0025	0.0125	0.025	0.25	25	2500	2.5×10^4	2.5×10^5	2.5×10^7
.0025	110	16	6	0.6 (54)	0.2 (7)	0.6 (15)	0.7 (17)	0.7 (17)	2 (15)
0.25	116	20	8	1 (44)	0.3 (5)	0.7 (15)	1 (17)	6 (22)	11 (39)
2.5	136	32	14	3 (52)	3 (8)	9 (23)	10 (25)	14 (29)	25 (67)
25	90	39	21	6 (56)	3 (8)	3 (21)	3 (24)	3 (25)	32 (82)
250	87	64	55	55 (69)	50 (53)	39 (87)	40 (96)	40 (97)	50 (116)
2500	133	98	93	101 (94)	91 (96)	69 (138)	71 (152)	71 (155)	71 (155)

$10^4 \times$ Average Variance and $10^3 \times$ Average Squared-Bias of $\hat{U}_2$ for Values of ϵ and δ on a Grid with $\epsilon = 0$. Values listed in parenthesis are for $\delta = 0.15$, but are not reported for values of $\delta < .25$, which undersmoothed in t when the data curves were contaminated with noise.

Table 1b

	$10^4 \times$ Variance	0.0025	0.0125	0.025	0.25	2500	2.5×10^4	2.5×10^5	2.5×10^7
0.0025	42	18	10	1 (426)	0.4 (44)	0.1 (32)	0.1 (23)	0.1 (14)	1 (8)
0.25	19	11	7	2 (151)	0.8 (18)	0.4 (23)	0.7 (23)	3 (19)	1 (6)
2.5	34	23	16	8 (133)	6 (19)	6 (17)	6 (18)	9 (21)	3 (9)
25	41	38	28	11 (163)	7 (28)	6 (20)	6 (21)	6 (20)	4 (11)
250	79	56	39	10 (199)	5 (30)	5 (14)	5 (14)	5 (14)	5 (16)
2500	122	56	37	10 (240)	7 (45)	7 (25)	8 (25)	8 (25)	8 (25)
	$10^3 \times$ Squared-Bias	0.0025	0.0125	0.025	25	2500	2.5×10^4	2.5×10^5	2.5×10^7
0.0025	27	10	6	0.8 (125)	0.2 (20)	0.1 (12)	0.1 (13)	0.1 (14)	3 (3)
0.25	25	10	5	0.7 (155)	0.2 (27)	0.2 (16)	0.3 (17)	3 (32)	8 (7)
2.5	48	26	16	5 (231)	5 (50)	5 (39)	5 (39)	7 (43)	12 (14)
25	39	26	16	2 (278)	0.5 (39)	0.2 (11)	0.2 (10)	0.3 (9)	13 (17)
250	39	24	18	8 (276)	9 (43)	14 (21)	15 (20)	15 (21)	17 (25)
2500	42	29	25	18 (332)	18 (69)	20 (34)	21 (33)	21 (33)	21 (34)