FUNCTIONAL DATA ANALYSIS

Functional data analysis (FDA) combines a variety of older methods and problems with some new perspectives, challenges and techniques for the analysis of data or models that involve functions. The main reference is [5], and [8] is a supplementary set of case studies.

Functional Data

The first panel of Figure 1 illustrates a number of aspects of functional data, as well as particular challenges to be taken up later. Functional data such as these ten records of the heights of children involve n discrete observations h_j, possibly subject to measurement error, that are assumed to reflect some underlying and usually smooth process. The argument values t_j associated with these values may not be equally spaced, and may not be the same from one record to another. We wish to consider any record, consisting in this case of the height measurements available for a specific child, as a single functional datum.

In the figure the arguments are values of time, but functional data may also be distributed over one or more spatial dimensions when the observation is an image, over both space and time, or may be taken from any other continuum. We usually assume that argument values are sampled from a continuum, even though the actual values are discrete.

Although the number n of discrete observations determines what may be achieved with functional data, the *data resolution* is more critical. This is defined as the smallest feature in the function that can be adequately defined by the data, where a curve feature is a peak or a valley, a crossing of a threshold, or some other localized shape characteristic of interest. The number of values of t_j per feature in the curve that are required to identify that feature depends on the nature of the feature and the amount of noise in the data. For example, with errorless data, at least three observations are required to identify the location, width, and amplitude of a peak; but if there is noise in the data, many more may be needed.

Functional Parameters

A functional data analysis may also involve functional parameters. Perhaps the example best known to statisticians is the probability density function $p(x)$ describing the distribution of a random variable x. If the density function is determined by a small fixed number of parameters, as with the normal density, then the model is parametric. But if we view the density simply as a function to be estimated from data without the imposition of any shape constraints except for positivity and possibly smoothness, then the density is functional parameter, and model for the data can be called functional, as opposed to parametric.

The smooth curves $h(t)$ fitting the height data in the left panel of Figure 1 are functional parameters for the discrete data if we estimate these ten curves in a way that is sufficiently open-ended to capture any detail in a curve that we need. In this growth model, we have an additional consideration: A curve h_i should logically be strictly increasing as well as smooth, and we will see below how to achieve this.

Functional models such as density estimates and smoothing curves are often called *nonparametric*, an unfortunate term because parameters are usually involved in the process, because there is a much older and mostly inconsistent use of this term in statistics, and because it is seldom helpful to describe something in terms of what it isn't.

We can also have functional parameters for nonfunctional data. Examples are intensity functions for point processes, item response functions used to model psychometric data, hazard functions for survival data, as well as probability density functions.

Some Roles for Derivatives

Functional data and parameters are usually assumed to be, at least implicitly, smooth or regular, and we consider why this should be so below. This implies in practice that a certain number of derivatives exist, and that these derivatives are sufficiently smooth that we can hope to use them in some way. We use the notation $D^m x(t)$ to indicate the value of

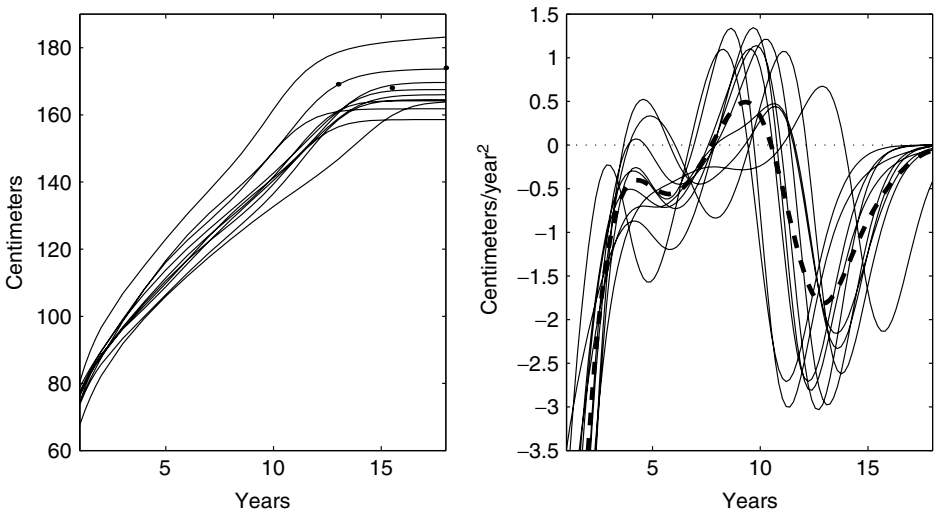

Figure 1. The left panel shows irregularly spaced observations of the heights of ten girls, along with the smooth increasing curves fit to each set of data. The right panel shows the second derivative or height acceleration functions computed from the smooth curves in the left panel. The heavy dashed line shows the mean of these acceleration curves.

the mth derivative of function x at argument value t.

The right panel of Figure 1 shows the estimated second derivative $D^2h(t)$ or acceleration of height for each girl. Height acceleration displays more directly than height itself the dynamics of growth in terms of the input of hormonal stimulation and other factors into these biological systems, just as acceleration in mechanics reflects exogenous forces acting on a moving body.

We also use derivatives to impose smoothness or regularity on estimated functions by controlling a measure of the size of a derivative. For example, it is common practice to control the size of the second derivative or curvature of a function using the total curvature measure $\int [D^2x(t)]^2\,dx$. The closer to zero this measure is, the more like a straight line the curve will be. More sophisticated measures are both possible and of genuine practical value.

A differential equation is an equation involving one or more derivatives as well as possibly the function value. For example, the differential equation defining harmonic or sinusoidal behavior is $\omega x + D^2x = 0$ for some positive constant ω. Differential equations can define a much wider variety of function behaviors than conventional parameter-based approaches, and this is especially so for nonlinear differential equations. Hence differential equations offer potentially powerful modelling strategies.

For example, returning to the need for a height function h to be strictly increasing or monotonic, it can be shown that any such function satisfies the differential equation

$$w(t)Dh(t) + D^2h(t) = 0, \qquad (1)$$

where $w(t)$ is some function that is itself unconstrained in any way. This equation transforms the difficult problem of estimating $h(t)$ to the much easier one of estimating $w(t)$, and also plays a role in curve registration discussed below. See [6] for other examples of differential equation models for functional parameters in statistics.

Finally, given actual functional data, can we estimate a differential equation whose solution will approximate the data closely? The technique *principal differential analysis* (PDA) for doing this can be an important part of a functional data analyst's toolbox. See the entry for this topic.

Historical Perspectives

Proposing a specific beginning to almost any field is controversial, but surely an early FDA landmark was Fourier's *Théorie de la Chaleur* (1822) demonstrating a powerful and computationally convenient basis function expansion system whose range of applications exceeds even today those of any other basis system for functional approximation. Techniques for analyzing certain special types of functional data have also been around for a long time. Time series analysis focusses on functions of time, where in practice time values are typically discrete and equally spaced. Assumptions of stationarity play a large role in this field, but tend to be avoided in FDA. Longitudinal data analysis is typically concerned with large numbers of replications of rather short time series, where the data in each record typically have little resolving power.

The term "functional data analysis" is rather recent, perhaps first used in [7]. At this early stage, the objective tended to be the adaptation of familiar multivariate techniques such as principal components analysis to functional contexts, but more recent work has extended the scope to include methods such principal differential analysis that are purely functional in nature.

SOME PERSPECTIVES ON FUNCTIONS

A function is a mapping from a set of objects called the *domain* of the function to another set of objects called its *range*, with the special property that any domain object is mapped to at most one range object. The natures of both the domain and range can be far wider than vectors of real or complex numbers. Indeed, mappings from a space of functions to another space of functions is especially important in this as in many fields, and we can indeed extend concepts like derivatives and integrals to these situations.

Whatever the domain and range, an essential property of a function is its *dimensionality*. In the majority of situations, functions can be viewed as points in an infinite dimensional space of possible functions, especially when the domain and/or the range is a continuum of some sort. Consequently, functions can be large in two quite independent ways: First, by how large their values are in the range space, and secondly, in terms of their dimensionality. White noise, for example, is an infinitely large object even if the noise is everywhere small. This presents a serious conceptual problem for the statistician: How can we hope to accurately estimate an infinite dimensional object from a finite amount of data?

But dimensionality, even if infinite, is not the same thing as *complexity*. Complexity has to do with the amount of detailed variation that a function exhibits over a small circumscribed region in the domain space. White noise, for example, is infinitely complex, but any reasonable account of a single child's growth curve should be fairly simple over a short time period. Complexity requires energy to achieve; and energy is everywhere in nature available in limited amounts, and can only be moved from one place to another at finite rates. The complexity of a stock market index, for example, although impressive, is still limited by the amount of capital available and the capacity of the financial community to move the capital around.

Mathematical tools for manipulating complexity and dimensionality independently are rather new. Multi-resolution analysis and wavelets offer basis systems in which the variation of a function can be split into layers of increasing complexity, but with the possibility of controlling the total energy allocated to each level. These concepts will surely have a large impact on the future of functional data analysis.

What model for ignorable variation, often called "noise" or "error", should we use? For finite dimensional data we tend to model noise as a set of N identically and independently distributed realizations of some random variable E. But the concept does not scale up in dimensionality well; white noise is infinite dimensional, infinitely complex, and infinitely large, and thus will ultimately overwhelm any amount of data. Moreover, putting the mathematical analysis of white noise and its close cousin, Brownian motion, in good order has challenged mathematicians for most of the last century. Ignorable functional variation, on the other hand, while admittedly more complex that we want to

deal with, is still subject to limits on the energy required to produce it. These issues are of practical importance when we consider hypothesis testing and other inferential issues.

FIRST STEPS IN FUNCTIONAL DATA ANALYSES

Techniques for Representing Functions

The first task in a functional data analysis is to capture the essential features of the data in a function. This we can always do by interpolating the data with a piecewise linear curve; this not only contains the original data within it, but also provides interpolating values between data points. However, some smoothing can be desirable as a means of reducing observational error, and if one or more derivatives are desired, then the fitted curve should also be differentiable at least up to the level desired, and perhaps a few derivatives beyond. Additional constraints may also be required such as positivity, monotonicity, and etc. The smoothing or regularization aspect of fitting can often be deferred, however, by applying it to the functional parameters to be estimated from the data rather than to the data-fitting curve itself.

Functions are traditionally defined by a small number of parameters that seem to capture the main shape features in the data, and ideally are also motivated by substantive knowledge about the process. For the height data, for example several models have been proposed, and the most satisfactory require eight or more parameters. But when the curve has a lot of detail or high accuracy is required, parametric methods usually break down. In this case two strategies dominate most of the applied work.

One approach is to use a system of basis functions $\phi_k(t)$ in the linear combination

$$x(t) = \sum_{k}^{K} c_k \phi_k(t). \qquad (2)$$

The basis system must have enough resolving power so that, with K large enough, any interesting feature in the data can be captured by x. Polynomials, where $\phi_k(t) = t^{k-1}$, have been more or less replaced by spline functions (see article for details) for nonperiodic data, but Fourier series remains the basis of choice for periodic problems. Both splines and Fourier series offer excellent control over derivative estimation. Wavelet bases are especially important for rendering sharp localized features.

In general, basis function expansions are easy to fit to data and easy to regularize. In the typical application the discrete observations y_j are fit by minimizing a penalized least squares criterion such as

$$\sum_{j}^{n} [y_j - x(t_j)]^2 + \lambda \int [D^2 x(t)]^2 \, dt \qquad (3)$$

When $x(t)$ is defined (2), solving for the coefficients c_k that minimize the criterion implies solving a matrix linear equation. The larger penalty parameter λ is, the more $x(t)$ will be like a straight line, and he close λ is to zero, the more nearly $x(t)$ will fit the data. There are a variety of data-dependent methods for finding an appropriate value of λ; see entry ?? for more details. It is also possible to substitute other more sophisticated differential operators for D^2, and consequently smooth toward targets other than straight lines; see [3] for examples and details.

For multi-dimensional arguments, the choice of bases become more limited. Radial and tensor-product spline bases are often used, but are problematical over regions with complex boundaries or regions where large areas do not contain observations. On the other hand, the triangular mesh and local linear bases associated with finite element methods for solving partial differential equations are extremely flexible, and software tools for manipulating them are widely available. Moreover, many data-fitting problems can be expressed directly as partial differential equation solutions; see [7] and [9] for examples.

The other function-fitting strategy is to use various types of kernel methods (see [4]), which are essentially convolutions of the data sequence with a specified kernel $K(s - t)$. The convolution of an image with a Gaussian kernel is a standard procedure in image analysis, and also corresponds to a solution of the heat or diffusion partial differential equation.

Basis function and convolution methods share many aspects, and each can be made to look like the other. They do, however, impose limits on what kind of function behavior can be accommodated in the sense that they cannot do more than is possible with the bases or kernels that are used. For example, there are situations where sines and cosines are, in a sense, too smooth, being infinitely differentiable, and for this reason are being replaced by wavelet bases in some applications, and especially in image compression where the image combines local sharp features with large expanses of slow variation.

Expressing both functional data and functional parameters as differential equations greatly expands the horizons of function representation. Even rather simple differential equations can define behavior that would be difficult to mimic in any basis system, and references on chaos theory and nonlinear dynamics offer plenty of examples; see [1]. Linear differential equations are already used in fields such as control theory to represent functional data, typically with constant coefficients. Linear equations with nonconstant coefficients are even more flexible tools, as we saw in equation (1 for growth curves, and nonlinear equations are even more so. More generally, functional equations of all sorts will probably emerge as models for functional data.

Registration and Descriptive Statistics

Once the data are represented by samples of functions, the next step might be to display some descriptive statistics. However, if we inspect a sample of curves such as the acceleration estimates in the right panel of Figure 1, we see that the timing of the pubertal growth spurt varies from child to child. The pointwise mean, indicated by the heavy dashed line, is a poor summary of the data; it has less amplitude variation and a longer pubertal growth period than those of any curve. The reason is that the average is taken over children doing different things at any point in time; some are pre-pubertal, some are in the middle, and some are terminating growth.

Registration is the one-to-one transformation of the argument domain in order to align salient curve or image features. That is, clock time is transformed for each child to biological time, so that all children are in the same growth phase with respect to the biological time scale. This transformation of time $h(t)$ is often called a *time warping function*, and it must, like a growth curve, be both smooth and strictly increasing. As a consequence, the differential equation (1) also defines $h(t)$, and, through the estimation of weight function $w(t)$, can be used to register a set of curves.

The mean function $\overline{x}(t)$ summarizes the location of a set of curves, possibly after registration. Variation among functions of a single argument is described by covariance and correlation surfaces, $v(s,t)$ and $r(s,t)$, respectively. These are bivariate functions that are the analogs of the corresponding matrices $\mathbf{V}$ and $\mathbf{R}$ in multivariate statistics. Of course, the simpler point-wise standard deviation curve $s(t)$ may also be of interest.

FUNCTIONAL DATA ANALYSES

FDA involves functional counterparts of multivariate methods such as principal components analysis, canonical correlation analysis, and various linear models. In principal components analysis, for example, the functional counterparts of the eigenvectors of a correlation matrix $\mathbf{R}$ are the eigenfunctions of a correlation function $r(s,t)$. Functional linear models use regression coefficient functions $\beta(s)$ or $\beta(s,t)$ rather than $\beta_1, \ldots, \beta_p$.

Naive application of standard estimation methods, however, can often result in parameters such as regression coefficient functions and eigenfunctions which are unacceptably rough or complex, even when the data are fairly smooth. Consequently, it can be essential to impose smoothness on the estimate, and this can be achieved through regularization methods, typically involving attaching a term to the fitting criterion that penalizes excessive roughness. Reference [5] offers a number of examples realized in the context of basis function representations of the parameters.

Few methods for hypothesis testing, interval estimation and other inferential problems in the context of FDA have been developed. Part of the problem is conceptual; what does

it mean to test a hypothesis about an infinite dimensional state of nature on the basis of finite information? On a technical level, although the theory of random functions or stochastic processes is well developed, testing procedures based on a white noise model for ignorable information are both difficult and seem unrealistic.

Much progress has been made in [10], however, in adapting t, F, and other tests to locating regions in multidimensional domains where the data indicate significant departures from a zero mean Gaussian random field. This works points the way to defining inference as a local decision problem rather than a global one.

Principal Components Analysis

Suppose that we have a sample of N curves $x_i(t)$, where the curves may have been registered and the mean function has been subtracted from each curve. Let $v(s,t)$ be their covariance function. The goal of principal components analysis (PCA) is to identify an important mode of variation among the curves, and we can formalize this as the search for a weighting function $\xi(t)$, called the *eigenfunction*, such that the amount of this variation, expressed as

$$\mu = \iint \xi(s)v(s,t)\xi(t)\,ds\,dt \tag{4}$$

is maximized subject to the normalizing condition $\int \xi^2(t)\,dt = 1$. This formulation of the PCA problem replaces the summations in the multivariate version of the problem, $\mu = v'\mathbf{V}v$, by integrals. The eigenequation, $\int v(s,t)\xi(t)\,dt = \mu\xi(s)$ defines the optimal solution, and subsequent principal components are defined as in the classic version; each eigenfunction $\xi_k(t)$ maximizes the sum of squared principal component scores subject to the normalizing condition and to the orthogonality condition $\int \xi_j(t)\xi_k(t)\,dt = 0, j < k$.

However, if the curves are not sufficiently smooth, the functional principal components defined by the $\xi_j(t)$ will tend to be unacceptably rough, and especially so for higher indices j. Smoothing the functions $x_i(t)$ will certainly help, but an alternate approach that may be applied to virtually all functional data analyses is to apply a roughness penalty directly to the estimated weight functions or eigenvalues $\xi(t)$. This is achieved by changing the normalizing condition to $\int \{\xi^2(t) + \lambda[D^2\xi(t)]^2\}\,dt = 1$. In this way the smoothing or regularization process, controlled by roughness penalty parameter λ, may be deferred to the step at which we estimate the functional parameter that interests us.

In multivariate work it is common practice to apply a rotation matrix $\mathbf{T}$ to principal components once a suitable number have been selected in order to aid interpretability. Matrix $\mathbf{T}$ is often required to maximize the VARIMAX criterion. The strategy of rotating functional principal components also frequently assists interpretation.

Linear Models

The possibilities for functional linear models are much wider than for the multivariate case. The simplest case arises when the dependent variable y is functional, the covariates are multivariate, and their values contained in a N by p design matrix $\mathbf{Z}$. Then the linear model is

$$\mathrm{E}[y_i(t)] = \sum_j^p z_{ij}\beta_j(t) \tag{5}$$

A functional version of the usual error sum of squares criterion is $\sum_i \int [y_i(t) - \hat{y}_i(t)]^2\,dt$, and the least squares solution for the regression coefficient functions $\beta_j(t)$ is simply $\hat{\beta}(t) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}(t)$ where $\mathbf{Y}(t)$ is the column vector of N dependent variable functions.

It will often happen that the covariate is functional, however. For example, we can have as the model for a scalar univariate observation y_i

$$\mathrm{E}[y_i] = \beta_0 + \int x_i(s)\beta(s)\,ds. \tag{6}$$

A functional linear model where both the dependent and covariate variables are functional that has been studied in some depth [2] is the varying coefficient or point-wise linear model involving p covariate functions

$$\mathrm{E}[y_i(t)] = \sum_j^p x_{ij}(t)\beta_j(t). \tag{7}$$

In this model, y_i depends on the x_{ij}'s only in terms of their simultaneous behavior.

But much more generally, a dependent variable function y_i can be fit by a single bivariate independent variable covariate x_i with the linear model

$$\mathrm{E}[y_i(t)] = \alpha(t) + \int_{\Omega_t} x(s,t)\beta(s,t)\,ds \qquad (8)$$

where $\alpha(t)$ is the intercept function, $\beta(s,t)$ is the bivariate regression function and Ω_t is a domain of integration that can depend on t. The notation and theorems of functional analysis can be used to derive the functional counterpart of a least squares fit of this model. In practice, moreover, we can expect that there will be multiple covariates, and some will be multivariate and some functional.

No matter what the linear model, we can attach one or more roughness penalties to the error sum of squares criterion being minimized to control the roughness the regression functions. Indeed, we *must* exercise this option when the dependent variable is either scalar or finite dimensional and the covariate is functional because the dimensionality of the covariate is then potentially infinite, and we can almost always find function a $\beta(s)$ that will provide a perfect fit. In this situation, controlling the roughness of the functional parameter also ensures that the solution is meaningful as well as unique.

Canonical Correlation Analysis

Canonical correlation analysis (CCA) in multivariate statistics is a technique for exploring covariation between two or more groups of variables. Although multivariate CCA is much less often used than PCA, its functional counterpart seems likely to see many applications because we often want to see what kind of variation is shared by two curves measured on the same unit, $x_i(t)$ and $y_i(t)$. Let $v_{XX}(s,t)$, $v_{YY}(s,t)$ and $v_{XY}(s,t)$ be the variance functions for the X variable, the Y variable, and the *covariance* function for the pair of variables, respectively. Here, we seek two *canonical weight* functions, $\xi(t)$ and $\eta(t)$, such that the *canonical correlation* criterion

$$\rho = \iint \xi(s)v_{XY}(s,t)\eta(t)\,ds\,dt \qquad (9)$$

is maximized subject to the two normalizing constraints

$$\iint \xi(s)v_{XX}(s,t)\xi(t)\,ds\,dt = 1$$

$$\iint \eta(s)v_{YY}(s,t)\eta(t)\,ds\,dt = 1.$$

Further matching pairs of canonical weight functions can also be computed by requiring that they be orthogonal to previously computed weight functions, as in PCA.

However, CCA is especially susceptible to picking up high frequency uninteresting variation in curves, and by augmenting the normalization conditions in the same way as for PCA by a roughness penalty, the canonical weight functions and the modes of covariation that they define can be kept interpretable. As in PCA, once a set of interesting modes of covariation have been selected, rotation can aid interpretation.

Principal Differential Analysis

Principal differential analysis (PDA) is a technique for estimating a linear variable-coefficient differential equation from one or more functional observations. That is, functional data are used to estimate an equation of the form

$$w_0(t)x(t) + w_1(t)Dx(t) + \ldots + w_{m-1}(t)D^{m-1}x(t) + D^m x(t) = f(t). \qquad (10)$$

The *order* of the equation is m, the highest order derivative used. The m weight coefficient functions $w_j(t)$ along with the *forcing function* $f(t)$ are what define the equation. Some of these functional parameters may be constant, and some may be zero. For example, the order two differential equation (1), defining growth and the time warping function $h(t)$ used in registration, has $w_0(t)$ and $f(t)$ equal to zero, and is therefore defined by the single weight function $w_1(t)$.

The main advantage of a differential equation model is that it captures the *dynamics* of the process in the sense it that also models velocity, acceleration and higher derivatives as well as the function itself. Moreover, differential equations of this sort, as well as *nonlinear* functions of derivatives, can define functional behaviors that are difficult

to capture in a low-dimensional basis function expansion.

See the entry for this topic for further details.

Acknowledgments

The author's work was prepared under a grant from the Natural Science and Engineering Research Council of Canada.

REFERENCES

1. Alligood, K. T., Sauer, T. D. and Yorke, J. A. (1996). *Chaos: An Introduction to Dynamical Systems.* Springer, New York, NY.
2. Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society*, Series B, **55**, 757–796.
3. Heckman, N. and Ramsay, J. O. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics*, **28**, 241–258.
4. Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Application—Theory and Methodologies.* Chapman and Hall, New York.
5. Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis.* Springer, New York, NY.
6. Ramsay, J. O. (2000). Differential equation models for statistical functions. *Canadian Journal of Statistics*, **28**, 225–240.
7. Ramsay, J.O. and Dalzell, C. (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 539–572.
8. Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis.* Springer, New York, NY.
9. Ramsay, T. (2002) Spline smoothing over difficult regions. *Journal of the Royal Statistical Society, Series B*, **64**, 307–319.
10. Worsley, K. J. (1994). Local maxima and the expected Euler characteristic of excusrsion sets of χ^2, F and t fields. *Advances in Applied Probability*, **26**, 13-42.

J. O. RAMSAY