The Future of Functional Data Analysis

Jiguo Cao (Simon Fraser University)
Jason Nielsen (Carleton University)
James Ramsay (McGill University)
Fang Yao (University of Toronto)

May 3–7, 2010

1 Overview of the Field

Functional data analysis concerns data providing information about curves, surfaces or anything else varying over a continuum. The continuum is often time, but may also be spatial location, wavelength, probability and etc.

The data may be so accurate that error can be ignored, may be subject to substantial measurement error, or even have a complex indirect relationship to the curve that they define. For example, measurements of the heights of children over a wide range of ages have an error level so small as to be ignorable for many purposes, but daily records of precipitation at a weather station are so variable as to require careful and sophisticated analyses in order to extract something like a mean precipitation curve.

However these curves are estimated, it is the assumption that they are intrinsically smooth that often defines a functional data analysis. In particular, functional data analyses often make use of the information in the slopes and curvatures of curves, as reflected in their derivatives. Plots of first and second derivatives, or plots of second derivative values as functions of first derivative values, may reveal important aspects of the processes generating the data. As a consequence, curve estimation methods designed to yield good derivative estimates can play a critical role in functional data analysis. Regularization is routinely employed to ensure smoothness in a derivative of a specified order, and also to quantify fidelity to a differential equation that may explain a substantial amount of the shape of the curve or surface.

Models for functional data and methods for their analysis may resemble those for conventional multivariate data, including linear and nonlinear regression models, principal components analysis, cluster analysis and most others. But the possibility of using derivative information greatly extends the power of these methods, and also leads to functional models defined by differential equations or dynamic systems, or other types of functional equations.

It has been clear from the beginning that curves and surfaces as data exhibit both phase and amplitude variation, where phase variation refers to the location on the continuous substrate of salient features in the curves. The first clear example of this was the temporal variation in the age of puberty in human growth curves, but subsequently phase variation became evident in many if not most samples of functional data. This has posed severe problems for the use of common descriptive statistics adapted to functional data, such as cross-sectional means, variances and correlations, as well as tools like principal components and regression analysis; all of which are designed to describe only amplitude variation. This bi–stochastic nature of functional data has since been recognized in many other branches of statistics, such as image analysis, shape analysis and tree-structured models.

The term "functional data analysis" was first used by [6], the first monograph was [3], and this was followed by [4] and [5]. [2] has subsequently appeared, and a number of other books are known to be in preparation.

As the workshop title indicates, the focus was less on surveying current and past research, and more on taking stock of where we've come, and then looking forward to anticipate the problems that we hope will inspire research in the coming years. We tried to divide the invitees to the workshop roughly evenly between the more senior members in the field who have already done much to define what FDA is today, and the young researchers with the potential to take this field to new places.

BIRS has moved this year to funding half workshops as well as the usual full workshop involving about 40 participants. We, as a half workshop, shared the facilities with another group over the Monday to Friday period of May 3 to 7. Our partner workshop was on Creative Writing in Mathematics and Science, and it would have been hard to choose a companion topic of more importance to the development of statistics. A number of us attended the Thursday evening session of the other workshop, and there was discussion of a more systematic interaction in the future.

Subtracting Wednesday afternoon, which by sacred tradition is given over to exploring the Rocky Mountains, this gave us nine morning/afternoon sessions of roughly three hours each, allowing for break time. We divided each of these in two, making 18 sessions of 1.5 hours each. This format gave us the opportunity to devote much more of the workshop to free unstructured exchanges than is typically the case, as well as making it possible for each of us to present our own work and exchange thoughts on the future of FDA. The amount and quality of the exchange was considered in our final evaluation to be perhaps the most important outcome of the week.

2 Recent Developments and Open Problems

We structured the week into themes:

- Random functions and inference and prediction
- Software, computational, numerical analysis and publication issues
- Estimating covariance structure, principal components analysis and functional variance components
- Statistical dynamics, both deterministic and stochastic
- Extension to spatial, spatial/temporal and other multidimensional domains
- Joint variation in amplitude and phase, the use of tensor methods
- Functional linear models, and input/output systems in general
- Native and observed coordinate and frame systems
- Applications

We also called attention to the forthcoming SAMSI Program on the Analysis of Object Oriented Data that aims to link functional data analysis, dynamic systems, shape analysis, image analysis and the analysis of tree-structured and other strongly non-Euclidean data. See http://www.samsi.info/programs/2010aoodprogram.shtml for more information.

3 Presentation Highlights

One of us (Ramsay) offered the following reflections.

When Bernard Silverman and I met in 1992 to write our first book, we knew a number of things. Our perspective on this emerging area would quickly be seen as too narrow. But a useful treatment of a restricted range of topics seemed much preferable to a scattered and disorganized account of everything that might come to mind. Keeping the math simple seemed paramount in order to maximize access to FDA methodology by

researchers with data to analyze. Although we did take a functional analytic approach in our own discussions, we knew the danger of even using the term "functional" in the title of the book, and we have heard so many times since that something deserving that qualifier must surely be too deep for ordinary people. As a consequence, we sacrificed depth in both mathematical and statistical terms to accessibility. The subsequent literature has done a fine job of providing much that we might have included and could not have provided due to our own limitations. Some advance was made in the 2005 edition, but much remains to be done.

But the workshop stretched the meaning of FDA far beyond what either of us could have envisaged, and Steve Marron's opening talk on object oriented data analysis was a tour de force of scene-setting in this sense. We learned from both Steve and Hans-Georg Müller that both the domains of functional data models and their range in some function space can have a manifold structure induced by a finite dimensional coordinate or chart system, which may or may not be local, that spans the actual variation in either of these spaces. This point was emphasized further by a number of applications as well as by the excellent discussion of the implications of "phase variation" and of the nature of a functional "feature".

The use of a dynamic system, either as a regularizer of a high-dimensional model, or as a model in its own right, also induces a manifold structure into the function space where the data are modeled. Both the null space of the associated differential operator and the variation in that null space induced by varying the parameters of the system seem important new aspects that we need to consider further. In addition to Hans-Georg's talk, that of Laura Sangalli also addressed directly the issue of how to estimate a manifold in model space. How do we estimate a space curve when there is no domain available except arc length, which of course only is defined by the estimate itself? And this in the presence of noisy data? The talk by Jianhua Huang on estimating the variation in boundaries of particles also seemed to fit into this manifold-structured data and model context.

Not nearly enough discussion was possible of extending the domain of functional data and models beyond one dimension to data distributed over space, space/time, and other multidimensional continua; but this seems surely a big topic for the time that we had available. We need another workshop on this alone, and a number of us are poised to extend FDA into spatial data analysis in the next couple of years.

But even in one-dimensional domains, we had a good deal of useful discussion of alternative measures of time that would be more appropriate to the data. Surjit Ray's presentation of the landsat data especially highlighted this issue. Debashis Paul's talk posed the question of how to work with intervals whose initial or final values are not known. It was recognized, too, that functional data often come as single or a small set of long series of observations having layers of structure, rather than as largish samples of "independent" functional observations, and that methods assuming replications, such as principal components analysis, need revisiting within this context. Simon Bonner's talk further developed this issue.

Bernard and I certainly did not appreciate how central the issue of the "right" coordinate system would become in FDA. Our first inkling of this was the appreciation of the need to estimate "system" time as opposed to clock time as a substrate for growth and weather data. Nevertheless, we too often used off-the-rack coordinate systems, such as orthogonal Cartesian coordinates for the handwriting and juggling data or latitude and longitude for spatial data, even when the data themselves clearly suggested better coordinate axes. Steve's "M-reps" as a boundary-defining method were especially striking. Diffusion-tensor imaging is also a recent approach to defining "intrinsic" coordinates for complex functional data. Triangulation methods using obvious feature-defined locations or cluster centers seems really natural in higher-dimensional settings.

It was inevitable that such a fascinating collection of data objects would inspire many comments on better ways to do functional data analysis. I can't do much better than listing a few of my favorites in point form.

- Neglecting auto-correlation over time or spatial covariation is a dangerous business, and that we did so little about this in both our books and in our software packages is embarrassing. This seems easy to correct, and we have to get at it.

- Methods like principal components analysis are essentially exploratory, and known components of variation such as mean effects, influences of obvious covariates like latitude and so forth, ought to be removed before using PCA and CCA on the residual structure. Otherwise we risk, or even will surely, mask interesting variation by using PCA to do the job that projections and regression methods were meant to do.

- We have to be careful with terminology. "Mean", "variance" and so forth are tightly tied to Hilbert space structures, and will mislead our collaborators when our models and analyses go beyond these

frameworks. Marc Genton's talk on displaying curve variation by functional box plots and Ivan Mizera's use of quantile regression seem just what we need in this regard. Finding better terminology might involve collaboration with the creative writing team that shared the BIRS facility with us.

- The issue of adding noise to models comes up every time I talk about dynamic systems. You all know now that this confuses me. I thought models were supposed to simplify the information in data, rather than simulating their complexity. Perhaps everyone should just give up on me.

- Outliers are a fact of life, and Liangliang Wang offered some radiosonde data that sure drove this point home, along with Ivan's emphasis on L1 based methodology. We need to improve our capacity to deal with this in the FDA toolbox.

I dove into the business of setting up an object-oriented FDA software package, first in Matlab and later in S-PLUS and R, with an enthusiasm that only can come with having no idea what one is getting into. Bernard warned me, but I refused to listen. Now I know, but at least I can say that people like Spencer Graves have come to my rescue in my worst moments, as well as those who wrote innumerable emails suggests corrections to errors and needed extensions.

Jason Nielsen's talk provided an exceptional overview of the positives and negatives of R and Matlab as software environments. He helped us all to understand why R is so slow, and how much faster it would run if it could be compiled. I can only say that we should all do a bit of fund-raising to give him the time he needs to finish his R compiler.

I've already mentioned tensor analysis as an essential tool as we get into manifolds and other aspects of differential geometry. How can we help our statistical colleagues to acquire this expertise with minimal effort? This is a question that has an analogue with respect to dynamic systems modeling. I've also mentioned the need to expand the FDA software to permit the modeling of auto- and spatial correlation, a simple task, it would seem.

Spatial and space-time FDA will require a rather more serious effort, but experience shows that there is no way around this task; if software is not readily available, they won't use it. In this respect, we seem stuck with the R environment for a long time to come. Basis function tools were commented on directly or indirectly many times. The use of what are called "empirical orthogonal functions" or "EOF's" in the physical science literature, but principal components by the rest of us, is now standard practice; and in my view a little too standard since it risks throwing away interesting variation. But it's here to stay and I'm extending the packages to allow for bases to be defined by eigenfunctions specifically and any functional data object in general. Also needed is the capacity to combine bases ($+$, $-$, and $*$ operators essentially) to allow for multilevel variation and other things. Jiguo and I [1] offer some tips in our paper on functional linear mixed modeling in the issue of JASA that has just appeared, and this will be in the next package releases. Not mentioned at the workshop but too important to omit here is the fact that Giles Hooker and a couple friends have released an R package CollocInfer for dynamic systems estimation along with a long manual.

Chunming Zhang was almost alone in considering the issue of inference for functional data, but in the balance this seems less surprising now than it did a couple of weeks ago. Inference is based on probability, and dare to question whether what is taught these days in courses on the subject will ever be of much help in this high-dimensional context. Perhaps probability theory is just low dimensional by its nature. How good it would be to be proven wrong about this!

4 Scientific Progress Made

Although the workshop could only bring together a small set of the rapidly expanding community of researchers and practitioners involved in functional data analysis, it did gather those who were exceptionally effective communicators and facilitators of discussion. Especially appreciated was the facilitation of involvement by new researchers in the discussion and the affirmation of their already significant achievements. The community development contribution of the workshop was therefore exceptional.

5 Outcome of the Meeting

The workshop will have a substantial impact on the SAMSI year-long project Analysis of Object Oriented Data. Many of the participants will also be involved in the opening SAMSI workshop in Sept. 12-15, 2010, and later on as organizers and researchers in residence.

The potential role of differential geometry in further developments in this field seemed obvious, and to suggest some hard work helping our colleagues to master tools such as tensor analysis. It was hoped that future workshops will bring together applied and pure mathematicians as well as statisticians in order to reflect in more depth on this theme.

The BIRS facility cannot be beat for its ambiance, which ensures delightful, leisurely and thoughtful discussion on a wide range of topics by participants coming to an area from many scientific domains. We particularly appreciated the warm hospitality and constant attention to supporting our work by Brenda Williams and her colleagues that were on site. The dining facilities at BIRS seemed like a week-long banquet, and the proximity of Banff town and Park, with their many opportunities for relaxation and exercise, contributed abundantly to the success of the workshop.

References

[1] J. Cao and J. O. Ramsay, Linear mixed effects modeling by parameter cascading. *Journal of the American Statistical Association*, **105**, 365–374, 2010.

[2] F. Ferraty and P. Vieu, *Nonparametric Functional Data Analysis*, Springer-Verlag, New York, 2006.

[3] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis* Springer-Verlag, New York, 1997.

[4] J. O. Ramsay and B. W. Silverman, *Applied Functional Data Analysis, Second Edition* Springer-Verlag, New York, 2002.

[5] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis, Second Edition* Springer-Verlag, New York, 2005.

[6] J. O. Ramsay and C. Dalzell, Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 365–380, 1991.

[7] Ramsay, J. O., Hooker, G., Cao, J. and Campbell, D. Parameter estimation for differential equations: A generalized smoothing approach (with discussion). *Journal of the Royal Statistical Society, Series B*. **69**, 741–796, 2007.