BIRS Workshop 15w5096

"Frontiers in Functional Data Analysis"

June 28 – July 3, 2015

Debashis Paul (University of California, Davis)
Surajit Ray (University of Glasgow)
David Ruppert (Cornell University)

1 Overview

The BIRS workshop "Frontiers in Functional Data Analysis" brought in about 40 experts in functional data analysis (FDA) and diverse scientific fields to assess the latest developments and explore in new directions in the field of FDA. There were 32 talks and three discussion sessions during the workshop focussing on both the current and future state of this rapidly evolving field. The talks were divided into the following main themes:

1. Methodological developments in FDA.
2. Dependent functional data and their applications.
3. Complex functional data objects.
4. Registration and dynamical systems view of functional data.
5. New approaches and new inferential techniques in FDA.
6. New applications of FDA.

A general consensus emerged regarding the future directions of FDA. First, the domain of applications of FDA is seen to be rapidly expanding, with emphasis on data with complex spatial, temporal and/or geometric structures. Secondly, the need for making rapid progress in developing computational tools for implementing methodologies as well as visualization tools is acknowledged. Concurrently, it is felt that mathematical challenges associated with the new methodological developments need to addressed and highlighted to attract a new generation of statisticians to this field. Finally, the participants saw a growing scope for a confluence of

FDA with spatio-temporal data analysis, even though historically these two fields developed independently of each other.

2 Recent developments

The workshop brought to fore the latest methodological developments in FDA. The workshop explored the developments in various directions, including (i) Bayesian formulation of FDA; (ii) structured functional regression models including mixed effects models and varying coefficients models; (iii) partial differential equation based regularization of functional data measured over smooth manifolds; (iv) functional time series and spatio-temporal FDA; (v) supervised functional principal component analysis; (vi) nonlinear representation of functional data through transformations; (vii) registration techniques for functional data; (viii) graphical models for functional data; (ix) representation of multi-way functional data; (x) regularization methods based on sparse representation; and (xi) visualization techniques for FDA.

In his inaugural keynote speech, David Ruppert gave an overview of functional dynamic linear models. He proposed a Bayesian formulation of the problem (Kowal, Matteson and Ruppert, 2014) and discussed its salient features, including, joint estimation of all the model parameters, exact inference on the parameters, and the flexibility to allow generalized dependences. He illustrated this framework through two data sets: (1) a local field potential data set on rats' brains where the goal is to understand the synchronization across different regions in the brain; and (2) data on yield curves for several economies. His talk underlined the considerable scope of modeling and inference provided by a Bayesian functional dynamic linear model for dealing with complex dependency relationships in functional data.

Sonja Greven gave an overview of functional linear regression models and clarified the distinct nature of scalar on function and function on scalar regression. She also gave a summary of methods dealing with covariates in functional regression framework with particular emphasis on functional additive mixed models (Scheipl, Staicu and Greven, 2015). Further, she emphasized the need for addressing issues such as robustness, interaction effects, variable selection, missingness, etc and gave a summary of various research projects dealing with these. She also illustrated different R-based software packages developed by her research group and collaborators. The latter include `FDboost` (Brockhaus, 2015), `mboost` (Hothorn et al., 2015), gamboostLSS (Hofner et al., 2015).

Jane-Ling Wang talked about a type of linear varying coefficient models for scalar on function regression with time-independent covariates (Zhang and Wang, 2015). Usually estimation of such models can be fairly complicated due to the presence of a multitude of parameters. She showed that a carefully specified identifiability constraint, followed by an integration in the time domain enables one to estimate the covariate effects separately and then to formulate a second stage to estimate the time dependent regression coefficient.

Among other notable developments in theoretical analysis of FDA techniques, Fang Yao formulated the problem of optimal recovery of functional data from noisy measurements within Le Cam's framework of asymptotic equivalence of experiments. He showed that under a Gaussian white noise assumption, a Stein shrinkage type adaptive estimation technique leads to optimal L^2-error in function reconstruction simultaneously over a class of n (sample size) Sobolev balls representing the underlying data space.

One of the traditional areas of application of FDA is chemometrics, especially using near infrared spectra (NIR) for prediction of, for example, octane of gasoline and percent fat of a meat sample. Frederic Ferraty discussed new tools for NIR data including functional projection pursuit and nonparametric variable selection (Ferraty et al., 2013). Siegfried Hörmann proposed a dynamic version of FPCA (functional principal components analysis) for general data structures (Hilbertian data) and studied its properties (Hörmann, Kidziński and Hallin, 2015). Alexander Aue proposed an approach to predicting functional autoregressive models through a combination of functional PCA and linear prediction based on Durbin-Watson and innovations algorithm and applied it to a particulate pollutant data (Aue, Dubart Norinho and Hörmann, 2015).

New methods for regularization of functional data have been presented by several participants, notably Philip Reiss, who developed principal coordinate ridge repression and regularization and Matthew Reimherr, who used group lasso based penalization for functional linear regression. Jiguo Cao focused on the problem of estimating parameters in varying coefficients models for PDEs and mixed effects models involving ODEs through parameter cascading approach (Wang et al., 2014).

Functions typically exhibit both phase and amplitude variation. Registration attempts to separate these two modes of variation. Although registration has been studied for many years, there was been substantial progress and new directions during recent years. Standard methods of registration align functions to shape features such as peaks and valley. Alois Kneip argues that doing this can be suboptimal. In his talk, he presented a registration method where warping functions are defined in such a way that the resulting registered curves span a low dimensional linear function space. Problems of identifiability were discussed in detail, and connections to established registration procedures were analyzed.

The need for representing data in an appropriate space is playing an increasingly important role in FDA. In his talk, Hans Müller focused on various approaches to transforming functional data before conducting downstream analysis. This included transformations using the average quantile function or the average hazard function with respect to the Wasserstein-Fréchet metric (Petersen and Müller, 2014). He showed its implications in separating the phase and amplitude variations in functional data. Further, he introduced a notion of functional manifold mean and functional manifold components (Chen and Müller, 2012).

3 Emerging fields of applications

Continuous monitoring mechanisms for various health and environmental indicators through widely accessible electronic devices has opened up a new frontier of functional data analysis. These data are typically noise-free and are measured over dense domains and therefore can be seen to have purely functional characteristics. One such data set on the sleeping pattern of individuals, based on actigraphy measurements, was analyzed by Jimin Ding within a functional mixed effects model framework.

Jeff Goldsmith's motivating example was a data set of binary curves indicating physical activity or inactivity are observed for nearly six hundred subjects over five days. He used a generalized linear model to incorporate scalar covariates into the mean structure, and decompose subject-specific and subject-day-specific deviations using multilevel functional principal components analysis. Model parameters were estimated in a Bayesian framework. In the application he identified effects of age and BMI on the time-specific change in probability of being active over a twenty-four hour period.

Some data can be modelled using a function whose form changes depending on some unobserved condition - that is, some unobserved state. Nancy Heckman investigated the energy used by a building on a given day and time with the outdoor temperature used as a covariate. This energy consumption depends on the number of coolers that are turned on, which is generally not observed in automatic energy monitoring systems (De Souza and Heckman, 2015). So the state is number of coolers that are on. Of interest is the expected energy consumption given the temperature and the number of coolers that are on. The probability of the number of coolers that are on is also of interest. In another example, the outcome is the three-dimensional acceleration of a marine mammal, with states "foraging" and "not foraging". Of interest is how and where and why a specified population of marine mammals feed.

Jianhua Huang developed methods for the analysis of two-way functional data using a two-way regularized singular value decompositions and illustrated this technique through an inverse problem of MEG source reconstruction (Tian et al., 2012). Kehui Chen provided theoretical support for an alternative approach to functional PCA for functional measurements over a multi-dimensional domain. Ana-Maria Staicu proposed a related technique for dealing with the problem of inference on fixed effects in functional and longitudinal data with illustration through Baltimore longitudinal study on aging (Park and Staicu, 2015). Haipeng Shen also considered a class of multi-level functional data with categorial covariates and proposed supervised principal components technique for representing such data through a combination of reduced rank regression and functional principal components (Li, Shen and Huang, 2015). He illustrated this method through a US financial company's call center arrival data.

Jeff Morris gave an overview of techniques for dealing with such multi-way and multi-factor functional data, including that of Bayesian modeling techniques – an avenue less explored within FDA paradigm.

Rob Hyndman presented new methods for exploring the feature space of large collections of time series. One of this related to dimension reduction and capturing seasonality information from multiple time series and the other on identifying anomalous time series, with illustration through web-traffic data. Göran Kauerman's talk was on the real time classification of fish in underwater sonar videos. He showed how after appropriate data preprocessing of the videos one can count and classify fish into different species based on their shape and movement. The developed procedures work in real time, that is data processing and classification of video sequences is faster than the length of the video sequences itself.

Climate data, e.g., the much-analyzed Canadian weather data set, has been a traditional area of application of FDA. Recently, new FDA methods for climate data have emerged. Piotr Kokoszka studied trends in the intensity of tropical storms. Applied to tropical storm data, the tests he introduced showed that there is a significant trend in the shape of the annual pattern of upper wind speed levels of hurricanes. Doug Nychka of the National Center for Atmospheric Research (NCAR) combined large-scale information from a global climate model with more detailed regional models to quantify uncertainty in regional predictions, for example, of extremes in daily precipitation. Surajit Ray studied an spatio temporal data involving vegetation pattern measured through remote sensing and showed how reduced rank separable spatio-temporal models can be used for analyzing such data (Liu, Ray and Hooker, 2014).

Much work has been done on the problem of detecting differences in the mean curves of two or more samples. Recently, researchers have tried to pinpoint exactly where in the functional domain they differ. In his talk, Jian Qing Shi developed methodology for the automatic detection of areas of significant differences using Gaussian process regression and controlling directional false discovery rates.

Marc Genton underlined the need for developing new visualization tools for FDA and demonstrated his research group's newly developed method of ranking functional data and its use in constructing functional box plots (Sun and Genton, 2011; Sun, Genton and Nychka, 2012) and animated visualization of functional data (Genton et al., 2015). He also emphasized the need for faster computational tools for implementing this and discussed some recent developments in this front.

Brain imaging has emerged has a data-rich area of study where FDA can prove to be a very useful tool. Hongtu Zhu and Owen Carmichael highlighted several potential applications within the field of brain imaging, including the issue of extracting information about the structure and functionality of the brain from functional MRI and diffusion tensor imaging data. These data can be viewed as functional data with a temporal component and information such as brain connectivity and influence of genetic and environmental factors are some questions where an FDA approach can be used effectively. fMRI data also serves as a prominent example of an emerging area, that is network analysis with functional variables. Carmichael raised the various issues such as validation, registration (for longitudinal or cross sectional data), interpretability, dimensionality effect etc that plague the current methods of analyzing such data and emphasized on the need of developing suitable FDA methodologies for statistically coherent analysis of such data. In a related vein, Hongxiao Zhu's talk focused on a Bayesian formulation of the problem of network analysis with functional variables within the Gaussian graphical model framework (Zhu, Strawn and Dunson, 2014).

4 Future directions

Much of the FDA literature contains linear inferential methods, although there has been some work on nonlinear methods such as kernel functional regression (Ferraty and Vieu, 2006). Hans Müller's keynote address surveyed recent progress on nonlinear methods and possible future developments.

Regularization using ordinary differential equations has been standard. Laura Sangalli used partial differential regularization in the analysis of functional data with complex structures and dependencies, e.g., neuro and medical imaging data (Azzimonti, 2013; Sangalli, Ramsay and Ramsay, 2013).

In his keynote address, Jim Ramsay showed how to reduce bias via regularization using an estimated differential operator that represents a model for the data. The orthogonal system defined by the operator is an effective method for representing the data.

FDA and graphical models are currently areas of intense research. Hongxiao Zhu's talk which connected them suggests a promising area for future work. She introduced a notion of conditional independence between random functions, and construct a framework for Bayesian inference of undirected, decomposable graphs in the multivariate functional data context.

During subsequent discussions it was identified that researchers in FDA need to collaborate and exchange ideas with each other more regularly. It was felt that one of the key areas where FDA can make significant contribution is spatio-temporal data analysis. This area is rich with data coming from numerous scientific applications, including climate studies, space weather, environmental science and ecology. With increasing progress in computational front, it was felt by the participants that having joint workshops with researchers from spatial statistics in near future will speed up collaboration among these two communities and exchange of ideas and tools will enrich scientific investigations. Participants also identified other potential areas of research which can contribute significantly to progress in FDA research. Thus joint workshops with mathematical modelers, especially people using PDEs and specific field of applications such as neuroscience was considered to be beneficial.

Another major focal point of the discussions was the increasing diversity of functional data and the increasingly prominent role of geometry in analyzing such data. These include data on non-euclidean surfaces such as the network of arteries, cortical surfaces and other anatomical objects. New geometrically adapted techniques for implementing smoothing and regression techniques on such surfaces are being developed as demonstrated through different talks during this workshop. With such abundance of data and methodologies, it was felt that providing a collection of well-catalogued data sets will go a long way in helping researchers in FDA to explore the capabilities of their methodologies. In addition, this will be useful in attracting talented young researchers to this field of study and developing new collaborations. Accordingly, it was proposed that the FDA community sets up a central repository of functional data sets and available analytical tools.

References

1. Aue, A., Dubart Norinho, D. and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, **110**, 319–348.

2. Azzimonti, L. (2013). *Blood flow velocity field estimation via spatial regression with PDE penalization.* Ph.D. Thesis, Politecnico Di Milano.

3. Brockhaus, S. Fuest, A., Mayr, A. and Greven, S. (2015). Functional regression models for location, scale and shape applied to stock return. In Friedl, H. and Wagner, H., (eds), *Proceedings of the 30th International Workshop on Statistical Modelling*, 117122.

4. Brockhaus, S., Scheipl, F., Hothorn, T. and Greven, S. (2015). The functional linear array model. *Statistical Modeling*, **15**, 279–300.

5. Chen, D. and Müller, H.-G. (2012). Nonlinear representation for functional data. *The Annals of Statistics*,

6. Crainiceanu, C. M., Staicu, A.-M. and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, **104**, 1550–1561.

7. de Souza, C. P. E. and Heckman, N. E. (2015). Modelling multi-level power usage with latent states and smooth functions. *arXiv:1504.02813*.

8. Ferraty, F., Goia, A., Salinelli, E. and Vieu, P. (2013). Functional projection pursuit regression. *Test*, **22**, 293–320.

9. Ferraty, F., and Vieu, P. (2006) *Nonparametric Functional Data Analysis*. Springer.

10. Genton, M. G., Castruccio, S., Crippa, P., Dutta, S., Huser, R., Sun, Y. and Vettori, S. (2015). Visuanimation in statistics. *Stat*, **4**, 81–96.

11. Hörmann, S., Kidziński, L. and Hallin, M. (2015). Dynamic functional principal components. *Journal of Royal Statistical Society, Series B*, **77**, 319–348.

12. Hofner, B., Mayr, A., Schmid, M. (2015). `gamboostLSS`: An R package for model building and variable selection in the GAMLSS framework". *Journal of Statistical Software* (to appear). *arXiv:1407.1774*.

13. Kowal, D. R., Matteson, D. S. and Ruppert, D. (2014). A Bayesian multivariate functional dynamic linear model. *arXiv:1411.0764*.

14. Li, G., Shen, H. and Huang, J. Z. (2015). Supervised sparse and functional principal component analysis. To appear in *Journal of Computational and Graphical Statistics*.

15. Liu, C., Ray, S. and Hooker, G. (2014). Functional principal components analysis of spatially correlated data. *arXiv:1411.4681*.

16. Park, S. Y. and Staicu, A.-M. (2015). Longitudinal functional data analysis. *arXiv:1506.08796*.

17. Petersen, A. and Müller, H.-G. (2014). Functional data analysis for density functions by transformation to a Hilbert space. *Technical report*.

18. Sangalli, L. M., Ramsay, J. O. and Ramsay, T. (2013). Spatial spline regression models. *Journal of Royal Statistical Society, Series B*, **75**, 1–23.

19. Scheipl, F., Staicu, A.-M. and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics* (to appear).

20. Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, **20**, 313-334.

21. Sun, Y., Genton, M. G. and Nychka, D. W. (2012). Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked ? *Stat*, **1**, 68–74.

22. Tian, T. S., Huang, J. Z., Shen, H. and Li, Z. (2012). A two-way regularization method for meg source reconstruction. *The Annals of Applied Statistics*, **6**, 1021–1046.

23. Wang, L., Cao, J., Ramsay, J. O., Burger, D., Laporte, C. and Rockstrohk, J. (2014). Estimating mixed-effects differential equation models. *Statistics and Computing*, **24**, 111–121.

24. Zhang, X. and Wang, J.-L. (2015). Varying-coefficient additive models for functional data. *Biometrika*, **102**, 15–32.

25. Zhu, H., Strawn, N. and Dunson, D. B. (2014). Bayesian graphical models for functional data. *arXiv:1411.4158*.