Guest Editorial for the Special Issue on Functional and Object Data Analysis

Hans-Georg Müller*
Department of Statistics, University of California, Davis
Davis, CA 95616, USA

December 2021

Functional data analysis (FDA) is a nonparametric statistical methodology for the modeling and analysis of data that include observations of random functions or data that are generated by latent random functions. It aims to impose as few parametric constraints as necessary, letting the data speak for themselves to the largest extent possible and to provide both exploratory and inferential tools for curve data and also longitudinal data. The field of FDA has much evolved since its theoretical beginnings that included the theoretical concept of functional regression models for Gaussian processes in Grenander (1950) and functional principal component analysis (FPCA) by Kleffe (1973) with extensions by Dauxois et al. (1982), emphasizing the perspective of data in Hilbert space.

Functional regression, basis expansions and FPCA to achieve dimension reduction for the inherently infinite-dimensional functional data remain the key statistical techniques of FDA; inference in these settings has been a more recent development. A modern introduction to the theoretical foundations of FDA with background on operators, reproducing kernel Hilbert spaces (RKHS) and random elements in Hilbert space can be found in Hsing and Eubank (2015). Major contributions to the theory were made by the late Peter Hall and his coauthors (Cai and Hall 2006; Hall et al. 2006; Hall and Horowitz 2007; Delaigle and Hall 2012). In addition to the theoretical developments, methodology designed for applications and accompanying software packages (mostly in R and Matlab) have broadened the appeal of FDA considerably. Applied FDA is now a well-established branch of data oriented nonparametric statistics, and offers numerous visualizations and exploratory tools for functional data, such as functional modes of variation.

FDA methodology evolved in the context of growth curves that provided motivation to develop techniques for nonparametric regression and targeting derivatives and time warping (especially via landmarks) in Theo Gasser's group (Gasser et al. 1984; Kneip

*This work was partially supported by National Science Foundation grant DMS-2014626.

and Gasser 1992) and through the extension of multivariate and psychometric methods such as PCA by Jim Ramsay (who coined the term "functional data analysis") and collaborators (Ramsay and Dalzell 1991; Ramsay and Silverman 2005). While a large part of FDA is based on an extension of multivariate data analysis methodology where matrices are replaced with linear operators and inner products are interpreted in infinite-dimensional Hilbert space (Eubank and Hsing 2008; He et al. 2018), its inherent analytical and stochastic aspects entail substantial challenges (Delaigle and Hall 2010; Koudstaal and Yao 2018; Lin et al. 2018).

Other developments that broadened the appeal and applicability of FDA include the sustained development of functional regression (Morris 2015); the incorporation of time warping methods (Gasser and Kneip 1995; Marron et al. 2015), where one considers additional random time distortions and is confronted with identifiability problems; the bridge to longitudinal data, which makes it possible to use FDA methodology for this large class of data (Staniswalis and Lee 1998; Rice and Wu 2001; Yao et al. 2005; Li and Hsing 2010; Zhang and Wang 2016); the connections with dynamics learning, empirical dynamics and fitting of differential equations (Ramsay et al. 2007; Dubey and Müller 2021); the case of partially observed functional data such as snippet or fragment data (Delaigle and Hall 2013; Dawson and Müller 2018; Descary and Panaretos 2019; Lin and Wang 2021); multivariate functional data (Zhou et al. 2008; Chiou and Müller 2014; Happ and Greven 2018); manifold-valued functional data and manifold learning (Chen and Müller 2012; Dai and Müller 2018; Lin and Yao 2019); classification and clustering of functional data (Chiou and Li 2007; Delaigle and Hall 2012); optimal designs for functional and longitudinal data (Ji and Müller 2017); and functional time series analysis (Bosq 2000; Panaretos and Tavakoli 2013), to name a few; for a review of some of these subfields see Wang et al. (2016). FDA methodology has been applied across the board in many areas of science and medicine, economics and finance, internet data, all kinds of longitudinal studies, and recently to the modeling of Covid-19 cases and deaths (Carroll et al. 2020; Boschi et al. 2021).

The field is very rich not least because it sits at the complex interface of smoothing, multivariate analysis, functional analysis, stochastic processes, longitudinal data, random effects modeling and dynamics. It is currently moving towards the study of more complex functional structures, including spatial and other complex dependencies, including connections with networks, repeatedly observed functional data, more sophisticated functional regression models and the modeling of extremes. These recent developments, confronting complex functional structures and data challenges are well represented in the collection of articles for this special issue.

A largely unexplored direction with rich potential for future research will also be the interface of FDA and random objects, i.e., metric-space valued random variables (Müller 2016) such as networks, trees, distributions and covariance structures, with connections to object-oriented data analysis (Marron and Alonso 2014). This interface includes the study of object-valued functional data (Dubey and Müller 2020) and distributional data (Petersen et al. 2021); a useful regression tool for random objects is Fréchet regression for global and local fitting and its variants (Petersen and Müller

2019; Chen and Müller 2020).

What follows is a brief overview of the papers that were included in the collection of articles for this special issue. While functional data are characterized by data samples that contain functions (which may be fully observed, observed on a dense grid, or partially observed, often with additional errors in the measurements) and thus has a focus on modeling samples that contain realizations of random functions and the underlying stochastic processes, FDA methodology relies heavily on smoothing techniques, especially as observed functional data often are collected as noise-contaminated measurements. Relevant smoothing methods include the classical approaches of kernel smoothing and local weighted least squares, as well as spline smoothing, B- and P-splines.

Applying smoothing splines to noisy data requires to choose a suitable penalty function, and this choice will impact the smoothing result. Bayesian methodology to achieve this difficult choice is the theme of Zhang et al. (2021). A second popular smoothing method, local linear regression, is used in Lin et al. (2021) to obtain direct estimates of Sharpe ratio functions in financial econometrics, where the Sharpe ratio is a variance-adjusted measure of how much better an investment is compared to a risk-free investment. In related situations where one works with single realizations of a process, this process may be multivariate and it is then of interest to relate its components to each other. This problem is addressed in Liu et al. (2021b), where complex ordinary differential equations are fitted by a novel application of deep learning methods. A central task for many FDA procedures is the estimation of mean and covariance functions of the underlying stochastic process from available discrete and noisy measurements. To address this, Cheng and Chen (2021) propose a framelet method and characterize the phase transitions one encounters when moving from dense to sparse sampling designs.

To enable FDA methodology for the analysis of ubiquitous longitudinal data that often feature sparse and irregular temporal designs, the PACE approach has become popular. It requires the estimation of mean and covariance functions as studied in Cheng and Chen (2021), but also requires inversion of a covariance matrix (of usually low dimension and where one may to employ regularization). As an alternative Nie et al. (2021) propose the SOAP method, aiming at a more targeted orthonormal approximation.

Predictor selection for concurrent functional regression, which has become a central statistical tool for many important applications, for example in longitudinal brain imaging (Wang et al. 2018; Chen et al. 2021), is studied in Ghosal and Maity (2021). The classical linear concurrent model is extended to allow for nonlinear relations and predictor selection. For a functional regression model where vector predictors are paired with functional responses, a predictor selection method is also developed in Cai et al. (2021) based on the SCAD approach (Fan and Li 2001), including automatic tuning parameter selection, which is often critical for good practical performance. Also pairing vector predictors with functional responses, Mehrotra and Maity (2021) study multicollinearity in this setting, grouping highly correlated predictors together.

Related to functional regression, another well-studied problem is binary classification for functional data. Many of the approaches proceed by employing projections, for example on a number of functional principal component scores; it is well known that this often suboptimal, as these projections do not take the responses into account. An improved projection method for classification is proposed in Zhou and Sang (2021). A related problem is functional clustering, which can be enhanced in the presence of covariates, as demonstrated in Jiang et al. (2021).

The modeling of extremes in the presence of functional data is of interest and there are many open problems in this area. Approaches to fill this gap include a RKHS model that is proposed for modeling expectiles of scalar responses coupled with functional predictors in Liu et al. (2021a) and a quantile approach to model extremes in a partial functional regression model, adopting FPCA in conjunction with extrapolation tools from extreme value theory in Zhu et al. (2021).

The challenges posed by spatial and temporal dependence of functional data have led to the subfields of spatial FDA and functional time series analysis, which are at the interface of FDA, spatial data analysis and time series analysis. For spatially indexed temporal point processes, motivated by Chicago Divvy bike sharing data, log-linear models for the intensity functions of the point processes are introduced in Gervini (2021). Another article with a focus on FDA for spatial data is Hörmann et al. (2021), where the problem of testing for Gaussianity is studied; this is an important problem in FDA as many methods rely implicitly or explicit on Gaussian assumptions.

Two articles focus on functional time series, where Sidrow et al. (2021) develop hidden Markov models for functions that are densely observed in time, with a very interesting application to data on Orca whales. Functional versions of potentially nonstationary fractionally integrated time series are studied in Shang (2021).

Finally, the work of co-editors Jiguo Cao, Guang Cheng and Yehua Li in handling the submissions for this special issue is gratefully acknowledged.

References

Boschi, T., Di Iorio, J., Testa, L., Cremona, M. A. and Chiaromonte, F. (2021). Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy. *Scientific Reports* **11** 1–15.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Springer, New York.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34** 2159–2179.

Cai, X., Xue, L. and Cao, J. (2021). Robust estimation and variable selection for function-on-scalar regression. *Canadian Journal of Statistics* .

Carroll, C., Bhattacharjee, S., Chen, Y., Dubey, P., Fan, J., Gajardo, A., Zhou, X., Müller, H.-G. and Wang, J.-L. (2020). Time dynamics of COVID-19. *Scientific Reports* **10** 21040.

Chen, D. and Müller, H.-G. (2012). Nonlinear manifold representations for functional data. *Annals of Statistics* **40** 1–29.

Chen, Y., Dubey, P., Müller, H.-G., Bruchhage, M., Wang, J.-L. and Deoni, S. (2021). Modeling sparse longitudinal data in early neurodevelopment. *NeuroImage* **237** 118079.

Chen, Y. and Müller, H.-G. (2020). Uniform convergence of local Fréchet regression, with applications to locating extrema and time warping for metric-space valued trajectories. *arXiv preprint arXiv:2006.13548* .

Cheng, K. and Chen, D.-R. (2021). Adaptive estimation for functional data: using framelet block thresholding method. *Canadian Journal of Statistics* .

Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B* **69** 679–699.

Chiou, J.-M. and Müller, H.-G. (2014). Linear manifold modelling of multivariate functional data. *Journal of the Royal Statistical Society: Series B* **76** 605–626.

Dai, X. and Müller, H.-G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Annals of Statistics* **46** 3334–3361.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis* **12** 136–154.

Dawson, M. and Müller, H.-G. (2018). Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippet data. *Journal of the American Statistical Association* **113** 1612–1624.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *Annals of Statistics* **38** 1171–1193.

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B* **74** 267–286.

Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association* **108** 1269–1283.

Descary, M.-H. and Panaretos, V. M. (2019). Recovering covariance from functional fragments. *Biometrika* **106** 145–160.

Dubey, P. and Müller, H.-G. (2020). Functional models for time-varying random objects. *Journal of the Royal Statistical Society B (with discussion)* **82** 275–327.

Dubey, P. and Müller, H.-G. (2021). Modeling time-varying random objects and dynamic networks. *Journal of the Americal Statistical Association (accepted for publication)* .

Eubank, R. L. and Hsing, T. (2008). Canonical correlation for stochastic processes. *Stochastic Processes and their Applications* **118** 1634–1661.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Gasser, T. and Kneip, A. (1995). Searching for structure in curve samples. *Journal of the American Statistical Association* **90** 1179–1188.

Gasser, T., Müller, H.-G., Köhler, W., Molinari, L. and Prader, A. (1984). Nonparametric regression analysis of growth curves. *Annals of Statistics* **12** 210–229.

Gervini, D. (2021). Doubly stochastic models for spatio-temporal covariation of replicated point processes. *Canadian Journal of Statistics* .

Ghosal, R. and Maity, A. (2021). Variable selection in nonparametric functional concurrent regression. *Canadian Journal of Statistics* .

Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv för Matematik* **1** 195–277.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics* **35** 70–91.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics* **34** 1493–1517.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association* **113** 649–659.

He, G., Müller, H.-G. and Wang, J.-L. (2018). Extending correlation and regression from multivariate to functional data. In *Asymptotics in Statistics and Probability*. De Gruyter, 197–210.

Hörmann, S., Kokoszka, P. and Kuenzer, T. (2021). Testing normality of spatially indexed functional data. *Canadian Journal of Statistics* .

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Ji, H. and Müller, H.-G. (2017). Optimal designs for longitudinal and functional data. *Journal of the Royal Statistical Society: Series B* **79** 859–876.

Jiang, J., Lin, H., Peng, H., Fan, G.-Z. and Li, Y. (2021). Cluster analysis with regression of non-Gaussian functional data on covariates. *Canadian Journal of Statistics* .

Kleffe, J. (1973). Principal components of random variables with values in a separable Hilbert space. *Mathematische Operationsforschung und Statistik* **4** 391–406.

Kneip, A. and Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics* **20** 1266–1305.

Koudstaal, M. and Yao, F. (2018). From multiple Gaussian sequences to functional data and beyond: a Stein estimation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 319–342.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics* **38** 3321–3351.

Lin, H., Tong, T., Wang, Y., Wenchao, X. and Riquan, Z. (2021). Direct local linear estimation for Sharpe ratio function. *Canadian Journal of Statistics* .

Lin, Z., Müller, H.-G. and Yao, F. (2018). Mixture inner product spaces and their application to functional data analysis. *Annals of Statistics* **46** 370–400.

Lin, Z. and Wang, J.-L. (2021). Mean and covariance estimation for functional snippets. *Journal of the American Statistical Association, accepted for publication* **xxx** xx–xx.

Lin, Z. and Yao, F. (2019). Intrinsic Riemannian functional data analysis. *Annals of Statistics* **47** 3533–3577.

Liu, M., Pietrosanu, M., Liu, P., Jiang, B., Zhou, X. and Kong, L. (2021a). Reproducing kernel-based functional linear expectile regression. *Canadian Journal of Statistics* .

Liu, Y., Li, L. and Wang, X. (2021b). A nonlinear sparse neural ordinary differential equation model for multiple functional processes. *Canadian Journal of Statistics* .

Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal* **56** 732–753.

Marron, J. S., Ramsay, J. O., Sangalli, L. M. and Srivastava, A. (2015). Functional data analysis of amplitude and phase variation. *Statistical Science* **30** 468–484.

Mehrotra, S. and Maity, A. (2021). Simultaneous variable selection, clustering, and smoothing in function-on-scalar regression. *Canadian Journal of Statistics* .

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application* **2** 321–359.

Müller, H.-G. (2016). Peter Hall, Functional Data Analysis and Random Objects. *Annals of Statistics* **44** 1867–1887.

Nie, Y., Yang, Y. and Cao, J. (2021). Recovering the underlying trajectory from sparse and irregular longitudinal data. *Canadian Journal of Statistics* .

Panaretos, V. M. and Tavakoli, S. (2013). Fourier analysis of stationary time series in function space. *Annals of Statistics* **41** 568–603.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics* **47** 691–719.

Petersen, A., Zhang, C. and Kokoszka, P. (2021). Modeling probability density functions as data objects. *Econometrics and Statistics, to appear* .

Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B* **53** 539–572.

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

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd ed. Springer Series in Statistics, Springer, New York.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259.

Shang, H. L. (2021). Not all long-memory estimators are born equal: The case of nonstationary functional time series. *Canadian Journal of Statistics* .

Sidrow, E., Heckman, N., Fortune, S. M., Trites, A. W., Murphy, I. and Auger-Méthé, M. (2021). Modelling multi-scale state-switching functional data with hidden Markov models. *Canadian Journal of Statistics* .

Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **93** 1403–1418.

Wang, H., Zhong, P.-S., Cui, Y. and Li, Y. (2018). Unified empirical likelihood ratio tests for functional concurrent linear models and the phase transition from sparse to dense functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 343–364.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and its Application* **3** 257–295.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100** 577–590.

Zhang, C., Kokoszka, P. and Petersen, A. (2021). Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis, to appear* .

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics* **44** 2281–2321.

Zhou, L., Huang, J. and Carroll, R. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95** 601–619.

Zhou, Z. and Sang, P. (2021). Continuum centroid classifier for functional data. *Canadian Journal of Statistics* .

Zhu, H., Yehua, L., Liu, B., Yao, W. and Zhang, R. (2021). Extreme quantile estimation for partial functional linear regression models with heavy-tailed distributions. *Canadian Journal of Statistics* .