FUNCTIONAL DATA ANALYSIS

Hans-Georg Müller
Department of Statistics
University of California, Davis
One Shields Ave., Davis, CA 95616, USA.
e-mail: mueller@wald.ucdavis.edu

KEY WORDS: Autocovariance Operator, Clustering, Covariance Surface, Eigenfunction, Infinite-dimensional Data, Karhunen-Loève Representation, Longitudinal Data, Nonparametrics, Panel Data, Principal Component, Registration, Regression, Smoothing, Square Integrable Function, Stochastic Process, Time Course, Tracking, Warping.

1. Overview

Functional data analysis (FDA) refers to the statistical analysis of data samples consisting of random functions or surfaces, where each function is viewed as one sample element. Typically, the random functions contained in the sample are considered to be independent and to correspond to smooth realizations of an underlying stochastic process. FDA methodology then provides a statistical approach to the analysis of repeatedly observed stochastic processes or data generated by such processes. FDA differs from time series approaches, as the sampling design is very flexible, stationarity of the underlying process is not needed, and autoregressive-moving average models or similar time regression models play no role, except where the elements of such models are functions themselves.

FDA also differs from multivariate analysis, the area of statistics that deals with finite-dimensional random vectors, as functional data are inherently infinite-dimensional and smoothness often is a central assumption. Smoothness has no meaning for multivariate data analysis, which in contrast to FDA is permutation invariant. Even sparsely and irregularly observed longitudinal data can be analyzed with FDA methodology. FDA thus is useful for the analysis of longitudinal or otherwise sparsely sampled data. It is also a key methodology for the analysis of time course, image and tracking data.

The approaches and models of FDA are essentially nonparametric, allowing for flexible modeling. The statistical tools of FDA include smoothing, e.g., based on series expansions, penalized splines, or local polynomial smoothing, and functional principal component analysis. A distinction between smoothing methods and FDA is that smoothing is typically used in situations where one wishes to obtain an estimate for one non-random object (where objects here are functions or surfaces) from noisy observations, while FDA aims at the analysis of a sample of random objects, which may be assumed to be completely observed without noise or to be sparsely observed with noise; many scenarios of interest fall in between these extremes.

An important special situation arises when the underlying random processes generating the data are Gaussian processes, an assumption that is often invoked to justify linear procedures and to simplify methodology and theory. Functional data are ubiquitous and may for example involve samples of density functions (Kneip and Utikal, 2001), hazard functions, or behavioral tracking data. Application areas that have been emphasized in the statistical literature include growth curves (Rao, 1958; Gasser et al., 1984), econometrics and e-commerce (Ramsay and Ramsey, 2002; Jank and Shmueli, 2006), evolutionary biology (Kirkpatrick and Heckman, 1989; Izem and Kingsolver, 2005), and genetics and genomics (Opgen-Rhein and Strimmer, 2006; Müller et al., 2008). FDA also applies to panel data as considered in economics and other social sciences.

2. Methodology

Key FDA methods include functional principal component analysis (Castro et al., 1986; Rice and Silverman, 1991), warping and curve registration (Gervini and Gasser, 2004) and functional regression (Ramsay and Dalzell, 1991). Theoretical foundations and asymptotic methods of FDA are closely tied to perturbation theory of linear operators in Hilbert space (Dauxois et al., 1982; Bosq, 2000; Mas and Menneteau, 2003); a reproducing kernel Hilbert space approach has also been proposed (Eubank and Hsing, 2008), as well as Bayesian approaches (Telesca and Inoue, 2008). Finite sample implementations typically require to address ill-posed problems, emplying suitable regularization, which is often implemented by penalized least squares or penalized likelihood and by truncated series expansions. A broad overview of methods and applied aspects of FDA can be found in the textbook Ramsay and Silverman (2005) and some additional reviews are in Rice (2004); Zhao et al. (2004); Müller (2008).

The basic statistical methodologies of ANOVA, regression, correlation, classification and clustering that are available for scalar and vector data have spurred analogous developments for functional data. An additional aspect is that the time axis itself may be subject to random distortions and adequate functional models sometimes need to reflect such time-warping (also referred to as alignment or registration).

Another issue is that often the random trajectories are not directly observed. Instead, for each sample function one has available measurements on a time grid that may range from very dense to extremely sparse. Sparse and randomly distributed measurement times are frequently encountered in longitudinal studies. Additional contamination of the measurements of the trajectory levels by errors is also common. These situations require careful modeling of the relationship between the recorded observations and the assumed underlying functional trajectories (Rice and Wu, 2001; James and Sugar, 2003; Yao et al., 2005a).

Initial analysis of functional data includes exploratory plotting of the observed functions in a "spaghetti plot" to obtain an initial idea of functional shapes, to check for outliers and to identify potential "landmarks". Preprocessing may include outlier removal and registration to adjust for time-warping (Gasser and Kneip, 1995; Gervini and Gasser, 2004; Liu and Müller, 2004; James, 2007; Kneip and Ramsay, 2008).

3. Functional Principal Components

Basic objects in FDA are the mean function μ and the covariance function G. For square integrable random functions $X(t)$,

$$\mu(t) = E(Y(t)), \quad G(s,t) \quad = \quad \operatorname{cov}\{X(s), X(t)\}, \quad s,t \in \mathcal{T}, \tag{1}$$

with auto-covariance operator $(Af)(t) = \int_{\mathcal{T}} f(s)G(s,t)\,ds$. This linear operator of Hilbert-Schmidt type has orthonormal eigenfunctions ϕ_k, $k = 1, 2, \ldots$, with associated ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$, such that $A\,\phi_k = \lambda_k\,\phi_k$. The foundation for functional principal component analysis is the Karhunen-Loève representation of random functions (Karhunen, 1946; Grenander, 1950; Gikhman and Skorokhod, 1969) $X(t) = \mu(t) + \sum_{k=1}^{\infty} A_k\,\phi_k(t)$, where $A_k = \int_{\mathcal{T}}(Y(t) - \mu(t))\phi_k(t)\,dt$ are uncorrelated centered random variables with $\operatorname{var}(A_k) = \lambda_k$, referred to as functional principal components (FPCs).

Estimation of eigenfunctions, eigenvalues and of FPCs is a core objective of FDA. Various smoothing-based methods and applications for various sampling designs have been considered (Jones and Rice, 1992; Silverman, 1996; Staniswalis and Lee, 1998; Cardot, 2000; James et al., 2000; Paul and Peng, 2009). Estimators employing smoothing methods (local least squares or splines) have been developed for various sampling schemes (sparse, dense, with errors) to obtain a data-based version of the eigen-representation, where one regularizes by truncating at a finite number K of included components. The idea is to borrow strength from the entire sample of functions, rather than estimating each function separately. The functional data are then represented by the subject-specific vectors of score estimates $\hat{A}_k$, $k = 1, \ldots, K$, which can be used to represent individual trajectories and for subsequent statistical analysis (Yao et al., 2005a).

More adequate representations of functional data are sometimes obtained by fitting pre-specified fixed basis functions with random coefficients. In particular, B-splines (Sy et al., 1997), P-splines (Yao and Lee, 2006) and wavelets (Morris and Carroll, 2006) have been successfully applied. A general relation between mixed linear models and fitting functional models with basis expansion coefficients can be used to advantage for modeling and implementation of these approaches. In the theoretical analysis, one may distinguish between an essentially multivariate analysis, which results from assuming that the number of series terms is actually finite, leading to parametric rates of convergence, and an essentially functional approach. In the latter, the number of components is assumed to increase with sample size and this leads to "functional" rates of convergence that depend on the properties of underlying processes, such as decay and spacing of the eigenvalues of the autocovariance operator.

4. Functional Regression and Related Models

Functional regression models may include one or several functions among the predictors, responses, or both. For pairs (X, Y) with centered random predictor functions X and scalar responses Y, the linear model is

$$E(Y|X) = \int_{\mathcal{T}} (X(s) - \mu(s))\beta(s)\, ds.$$

The regression parameter function β can be represented in a suitable basis, for example the eigenbasis, with coefficient estimates determined by least squares or similar criteria. The functional linear model has been thoroughly studied, including optimal rates of convergence (Cardot et al., 2003a,b; Yao et al., 2005b; Cai and Hall, 2006; Hall and Horowitz, 2007; Li and Hsing, 2007; Mas and Pumo, 2009).

The class of useful functional regression models is large, due to the infinite-dimensional nature of the functional predictors. The case where the response is functional (Ramsay and Dalzell, 1991) also is of interest. Flexible extensions of the functional linear model for example include nonparametric approaches (Ferraty and Vieu, 2006), where unfavorable small ball probabilities and the non-existence of a density in general random function space impose limits on convergence (Hall et al., 2009), and multiple index models (James and Silverman, 2005). Another extension is the functional additive model (Müller and Yao, 2008). For functional predictors $X = \mu + \sum_{k=1}^{\infty} A_k \phi_k$ and scalar responses Y, this model is given by

$$E(Y|X) = \sum_{k=1}^{\infty} f_k(A_k)\phi_k$$

for smooth functions f_k with $E(f_k(A_k)) = 0$.

Another variant of the functional linear model, which is also applicable for classification purposes, is the generalized functional linear model $E(Y|X) = g\{\mu + \int_{\mathcal{T}} X(s)\beta(s)\,ds\}$ with link function g (James, 2002; Escabias et al., 2004; Cardot and Sarda, 2005; Müller and Stadtmüller, 2005). The link function (and an additional variance function if applicable) is adapted to the (often discrete) distribution of Y; the components of the model can be estimated by quasi-likelihood. Besides discriminant analysis via the binomial functional generalized linear model, various other methods have been studied for functional clustering and discriminant analysis (James and Sugar, 2003; Chiou and Li, 2007, 2008).

Of practical relevance are extensions towards polynomial functional regression models (Müller and Yao, 2010), hierarchical functional models (Crainiceanu et al., 2009), models with varying domains, models with more than one predictor function, and functional (autoregressive) time series models, among others. In addition to the functional trajectories themselves, derivatives are of interest to study the dynamics of the underlying processes (Ramsay and Silverman, 2005). Software for functional data analysis evolves rapidly and is available from various sources. Freely available software includes for example the fda package (R and matlab), at `http://www.psych.mcgill.ca/misc/fda/software.html`,
and the PACE package (matlab), at `http://anson.ucdavis.edu/ mueller/data/pace.html`.

Acknowledgments

Research supported in part by NSF Grant DMS-0806199. Based on an article from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC.

References

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

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

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics* **12** 503–538.

Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2003a). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics. Theory and Applications* **30** 241–255.

Cardot, H., Ferraty, F. and Sarda, P. (2003b). Spline estimators for the functional linear model. *Statistica Sinica* **13** 571–591.

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* **92** 24–41.

Castro, P. E., Lawton, W. H. and Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28** 329–337.

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

Chiou, J.-M. and Li, P.-L. (2008). Correlation-based functional clustering via subspace projection. *Journal of the American Statistical Association* **103** 1684–1692.

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

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.

Escabias, M., Aguilera, A. M. and Valderrama, M. J. (2004). Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics* **16** 365–384.

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

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

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. *The Annals of Statistics* **12** 210–229.

Gervini, D. and Gasser, T. (2004). Self-modeling warping functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 959–971.

Gikhman, I. I. and Skorokhod, A. V. (1969). *Introduction to the Theory of Random Processes.* W. B. Saunders Company, Philadelphia.

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. *The Annals of Statistics* **35** 70–91.

Hall, P., Müller, H.-G. and Yao, F. (2009). Estimation of functional derivatives. *The Annals of Statistics* **37** 3307–C3329.

Izem, R. and Kingsolver, J. (2005). Variation in continuous reaction norms: Quantifying directions of biological interest. *American Naturalist* **166** 277–289.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 411–432.

James, G. M. (2007). Curve alignment by moments. *Annals of Applied Statistics* **1** 480–501.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602.

James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association* **100** 565–576.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98** 397–408.

Jank, W. and Shmueli, G. (2006). Functional data analysis in electronic commerce research. *Statistical Science* **21** 155–166.

Jones, M. C. and Rice, J. A. (1992). Displaying the important features of large collections of similar curves. *The American Statistician* **46** 140–145.

Karhunen, K. (1946). Zur Spektraltheorie stochastischer Prozesse. *Annales Academiae Scientiarum Fennicae. Series A. I, Mathematica* **1946** 7.

Kirkpatrick, M. and Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27** 429–450.

Kneip, A. and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103** 1155–1165.

Kneip, A. and Utikal, K. J. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* **96** 519–542.

Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *Journal of Multivariate Analysis* **98** 1782–1804.

Liu, X. and Müller, H.-G. (2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association* **99** 687–699.

Mas, A. and Menneteau, L. (2003). Perturbation approach applied to the asymptotic study of random operators. In *High dimensional probability, III (Sandjberg, 2002)*, vol. 55 of *Progr. Probab.* Birkhäuser, Basel, 127–134.

Mas, A. and Pumo, B. (2009). Functional linear regression with derivatives. *Journal of Nonparametric Statistics* **21** 19–40.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 179–199.

Müller, H.-G. (2008). Functional modeling of longitudinal data. In *Longitudinal Data Analysis (Handbooks of Modern Statistical Methods)* (G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs, eds.). Chapman & Hall/CRC, New York, 223–252.

Müller, H.-G., Chiou, J.-M. and Leng, X. (2008). Inferring gene expression dynamics via functional regression analysis. *BMC Bioinformatics* **9** 60.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics* **33** 774–805.

Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103** 1534–1544.

Müller, H.-G. and Yao, F. (2010). Functional quadratic regression. *Biometrika* **97** 49–64.

Opgen-Rhein, R. and Strimmer, K. (2006). Inferring gene dependency networks from genomic longitudinal data: A functional data approach. *REVSTAT - Statistical Journal* **4** 53–65.

Paul, D. and Peng, J. (2009). Consistency of restricted maximum likelihood estimators in functional principal components analysis. *Annals of Applied Statistics* **37** 1229–1271.

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

Ramsay, J. O. and Ramsey, J. B. (2002). Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics* **107** 327–344. Information and entropy econometrics.

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

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics* **14** 1–17.

Rice, J. A. (2004). Functional and longitudinal data analysis: Perspectives on smoothing. *Statistica Sinica* 631–647.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **53** 233–243.

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

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* **24** 1–24.

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

Sy, J. P., Taylor, J. M. G. and Cumberland, W. G. (1997). A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics* **53** 542–555.

Telesca, D. and Inoue, L. Y. (2008). Bayesian hierarchical curve registration. *Journal of the American Statistical Association* **103** 328–339.

Yao, F. and Lee, T. C. M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 3–25.

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

Yao, F., Müller, H.-G. and Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33** 2873–2903.

Zhao, X., Marron, J. S. and Wells, M. T. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica* **14** 789–808.