

FUNCTIONAL DATA ANALYSIS

Contribution to the
International Handbook (Encyclopedia) of Statistical Sciences

July 28, 2009

Hans-Georg Müller¹

Department of Statistics

University of California, Davis

One Shields Ave., Davis, CA 95616, USA.

e-mail: mueller@wald.ucdavis.edu

¹Research partially supported by NSF Grant DMS-0806199

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 smooth. FDA methodology is essentially nonparametric, utilizes smoothing methods, and allows for flexible modeling. The underlying random processes generating the data are sometimes assumed to be (non-stationary) Gaussian processes.

Functional data are ubiquitous and may involve samples of density functions (Kneip and Utikal, 2001) or hazard functions (Chiou and Müller, 2009). Application areas include growth curves, econometrics, evolutionary biology, genetics and general kinds of longitudinal data. FDA methodology features functional principal component analysis (Rice and Silverman, 1991), warping and curve registration (Gervini and Gasser, 2004) and functional regression (Ramsay and Dalzell, 1991). Theoretical foundations and asymptotic analysis of FDA are closely tied to perturbation theory of linear operators in Hilbert space (Bosq, 2000). Finite sample implementations often require to address ill-posed problems with suitable regularization. A broad overview of applied aspects of FDA can be found in the textbook Ramsay and Silverman (2005).

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. 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., 2005). Initial analysis of functional data includes exploratory plotting of the observed functions in a “spaghetti plot” to obtain an initial idea of functional shapes, check for outliers and identify “landmarks”. Preprocessing may include

outlier removal and curve alignment (registration) to adjust for time-warping.

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

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

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 , $k = 1, 2, \dots$, with associated ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots$, such that $A\phi_k = \lambda_k \phi_k$. The foundation for functional principal component analysis is the Karhunen-Loève representation of random functions $X(t) = \mu(t) + \sum_{k=1}^{\infty} A_k \phi_k(t)$, where $A_k = \int_{\mathcal{T}} (Y(t) - \mu(t))\phi_k(t) dt$ are uncorrelated centered random variables with $\text{var}(A_k) = \lambda_k$.

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 this 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 $\hat{A}_k$, $k = 1, \dots, K$, which can be used to represent individual trajectories and for subsequent statistical analysis. Useful representations are alternatively obtained with pre-specified fixed basis functions, notably B-splines and wavelets.

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

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

The regression parameter function β is usually represented in a suitable basis, for example the eigenbasis, with coefficient estimates determined by least squares or similar criteria. A variant, 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 . 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.

The class of useful functional regression models is large. A flexible extension of the functional linear model is the functional additive model. Writing centered predictors as $X = \sum_{k=1}^{\infty} A_k \phi_k$, it is given by

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

for smooth functions f_k with $E(f_k(A_k)) = 0$. Of practical relevance are models with varying domains, with more than one predictor function, and functional (autoregressive) time series models. In addition to the functional trajectories themselves, their derivatives are of interest to study the dynamics of the underlying processes.

References

- BOSQ, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Springer-Verlag, New York.
- CHIOU, J.-M. and MÜLLER, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association* **104** 572–585.
- GERVINI, D. and GASSER, T. (2004). Self-modeling warping functions. *Journal of the Royal Statistical Society: Series B* **66** 959–971.
- JAMES, G. M. and SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98** 397–408.
- 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.
- 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. and SILVERMAN, B. W. (2005). *Functional Data Analysis*. 2nd ed. Springer Series in Statistics, Springer, New York.

- 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* **53** 233–243.
- RICE, J. A. and WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259.
- 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.