

Benchmarking Methods for Functional Data Analysis

Supervision: Marlène Baumeister and Markus Pauly, jointly with Łukasz Smaga and Natalia Stefanska (Adam Mickiewicz University, Poznań)

Research environment: International collaboration within the WEAVE project by Łukasz Smaga and Markus Pauly

Functional Data Analysis (FDA) provides statistical tools for observations that are naturally represented as curves or functions, such as longitudinal trajectories, biomedical signals, growth curves, spectroscopy measurements, wearable-sensor data, EEG data, or other measurements observed over time or space. A large variety of statistical methods has been proposed for such data, including methods for functional regression as well as inferential procedures for factorial designs with functional responses.

While new FDA methods are frequently evaluated in simulation studies or on a small number of illustrative datasets, much less is known about their comparative performance across a broad and heterogeneous collection of real-world datasets. In particular, the performance of a method usually depends strongly on sample sizes, dimensionality, covariance structure, sparsity, balance of the design, or the strength and type of functional effects.

The aim of this thesis is to develop a systematic benchmark study for FDA based on openly available datasets, with a particular focus on applications from the life sciences.

The project may consist of the following components:

1. Collection of open functional datasets.

Collect openly available datasets that are suitable for methodological comparison. Of particular interest are datasets arising in biological and life-science applications, such as spectroscopy, biomechanics, longitudinal measurements, neuroimaging, physiological signals, EEG, ecological curves (pollution, weather), or other high-resolution measurements.

The datasets should cover settings relevant for, e.g.,

- scalar-on-function, function-on-scalar, or function-on-function regression,
- one-way or multi-factor functional designs (univariate, multivariate or with repeated measurements),
- balanced and unbalanced designs,
- densely and sparsely observed functions.

2. **Systematic characterization of datasets.**

Characterize the datasets according to statistical properties that may affect method performance. Possible characteristics include

- sample size and number of groups,
- grid size and observation density,
- smoothness and signal-to-noise ratio,
- covariance and correlation structure,
- sparsity and irregular observation times,
- dimensionality and effective number of functional principal components,
- balance or imbalance of factorial designs,
- magnitude and shape of group or regression effects,
- missingness (partial or complete).

The resulting collection could itself form a useful benchmark repository for FDA.

3. **Large-scale benchmark comparison of FDA methods.**

Apply a selection of established and more recent methods to the curated datasets and compare their empirical behaviour.

Depending on the chosen focus, the study could investigate methods for functional regression, functional ANOVA/MANOVA and factorial designs, or both.

Relevant performance criteria may include predictive performance, stability, computation time, sensitivity to preprocessing or tuning choices, and if applicable, estimation accuracy or inferential behaviour.

4. **Optional extension: plasmode-based data-generating processes.**

A particularly interesting methodological extension would be to use selected real datasets as the basis for plasmode simulations (models). Instead of specifying the entire functional data-generating process parametrically, important aspects of the observed functional structure could be retained through resampling or model-based reconstruction, while known group or regression effects are imposed artificially.

This would combine the realism of empirical functional datasets with the main advantage of simulation studies: a partially known ground truth. Such plasmode data-generating mechanisms could then be used to study bias, coverage, type-I error, power, or prediction

error under data structures that more closely resemble real applications.

The exact scope can be adapted to the student's background and interests. A primarily empirical benchmark based on real datasets could be suitable for a B.Sc. or M.Sc. thesis, whereas the development and systematic evaluation of functional plasmode generators would make most sense within a M.Sc. thesis.

The project is particularly suitable for students interested in R/Python, reproducible research, statistical computing, functional data analysis, and method comparison.

Some References

- Febrero-Bande, M., Oviedo de la Fuente, M., Darbalaei, M., & Amini, M. (2022). Functional regression models with functional response: A new approach and a comparative study.
- Gertheiss, J., Goldsmith, J., & Staicu, A.-M. (2017). A note on functional regression. / Related comparative literature on scalar-on-function regression.
- Gertheiss, J., Rügamer, D., Liew, B. X., & Greven, S. (2024). Functional data analysis: An introduction and recent developments. *Biometrical Journal*, 66(7), e202300363.
- Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2, 321–359.
- Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer.
- Schreck, N., Slynko, A., Saadati, M., & Benner, A. (2024). Statistical plasmode simulations - Potentials, challenges and recommendations. *Statistics in Medicine*.