Functional Data Analysis

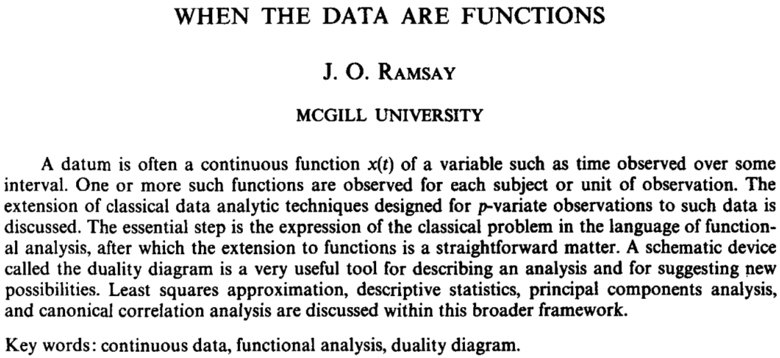
Whitney Huang

CLEMSON UNIVERSITY

wkhuang@clemson.edu

Clemson ENVR Group, November 4, 2020

Ramsay (1982), Psychometrika



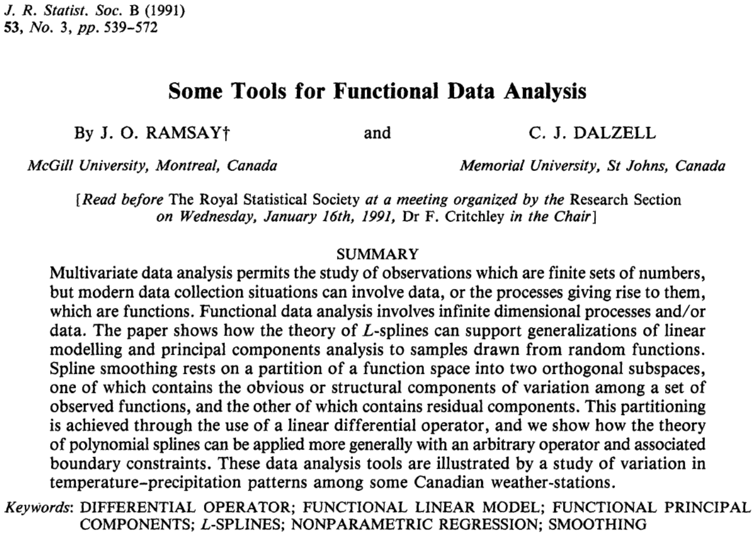
WHEN THE DATA ARE FUNCTIONS

J. O. Ramsay

MCGILL UNIVERSITY

A datum is often a continuous function $x(t)$ of a variable such as time observed over some interval. One or more such functions are observed for each subject or unit of observation. The extension of classical data analytic techniques designed for p-variate observations to such data is discussed. The essential step is the expression of the classical problem in the language of functional analysis, after which the extension to functions is a straightforward matter. A schematic device called the duality diagram is a very useful tool for describing an analysis and for suggesting new possibilities. Least squares approximation, descriptive statistics, principal components analysis, and canonical correlation analysis are discussed within this broader framework.

Key words: continuous data, functional analysis, duality diagram.



J. R. Statist. Soc. B (1991)
53, *No.* 3, *pp.* 539–572

Some Tools for Functional Data Analysis

By J. O. RAMSAY† and C. J. DALZELL

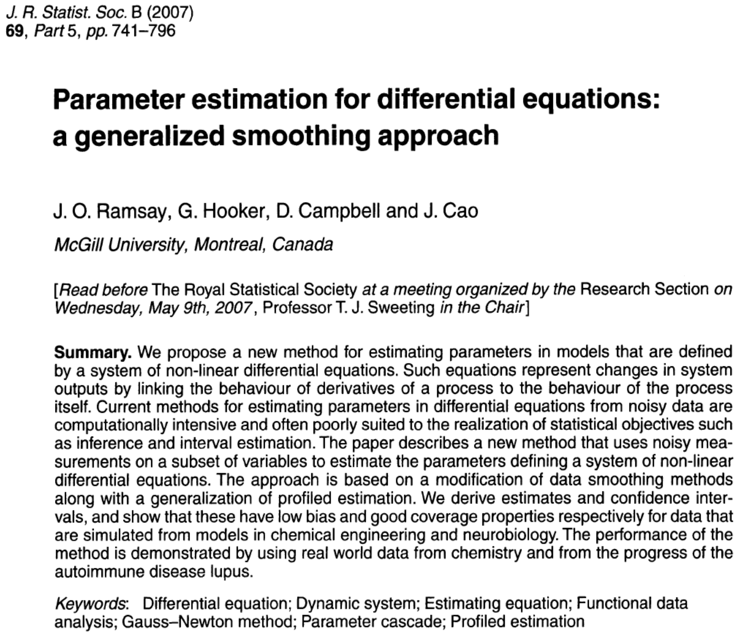
McGill University, Montreal, Canada — *Memorial University, St Johns, Canada*

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, January 16th, 1991,* Dr F. Critchley *in the Chair*]

SUMMARY

Multivariate data analysis permits the study of observations which are finite sets of numbers, but modern data collection situations can involve data, or the processes giving rise to them, which are functions. Functional data analysis involves infinite dimensional processes and/or data. The paper shows how the theory of *L*-splines can support generalizations of linear modelling and principal components analysis to samples drawn from random functions. Spline smoothing rests on a partition of a function space into two orthogonal subspaces, one of which contains the obvious or structural components of variation among a set of observed functions, and the other of which contains residual components. This partitioning is achieved through the use of a linear differential operator, and we show how the theory of polynomial splines can be applied more generally with an arbitrary operator and associated boundary constraints. These data analysis tools are illustrated by a study of variation in temperature–precipitation patterns among some Canadian weather-stations.

Keywords: DIFFERENTIAL OPERATOR; FUNCTIONAL LINEAR MODEL; FUNCTIONAL PRINCIPAL COMPONENTS; *L*-SPLINES; NONPARAMETRIC REGRESSION; SMOOTHING



J. R. Statist. Soc. B (2007)
69, *Part* 5, *pp.* 741–796

Parameter estimation for differential equations: a generalized smoothing approach

J. O. Ramsay, G. Hooker, D. Campbell and J. Cao

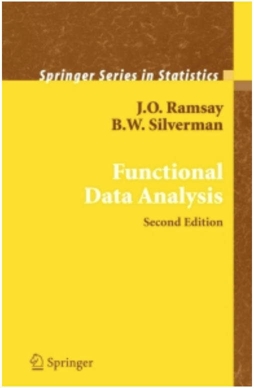
McGill University, Montreal, Canada

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, May 9th, 2007*, Professor T. J. Sweeting *in the Chair*]

Summary. We propose a new method for estimating parameters in models that are defined by a system of non-linear differential equations. Such equations represent changes in system outputs by linking the behaviour of derivatives of a process to the behaviour of the process itself. Current methods for estimating parameters in differential equations from noisy data are computationally intensive and often poorly suited to the realization of statistical objectives such as inference and interval estimation. The paper describes a new method that uses noisy measurements on a subset of variables to estimate the parameters defining a system of non-linear differential equations. The approach is based on a modification of data smoothing methods along with a generalization of profiled estimation. We derive estimates and confidence intervals, and show that these have low bias and good coverage properties respectively for data that are simulated from models in chemical engineering and neurobiology. The performance of the method is demonstrated by using real world data from chemistry and from the progress of the autoimmune disease lupus.

Keywords: Differential equation; Dynamic system; Estimating equation; Functional data analysis; Gauss–Newton method; Parameter cascade; Profiled estimation

FDA Books



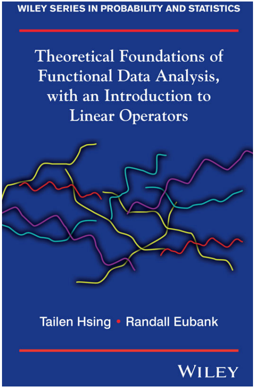
Springer Series in Statistics

J.O. Ramsay
B.W. Silverman

Functional Data Analysis

Second Edition

Springer

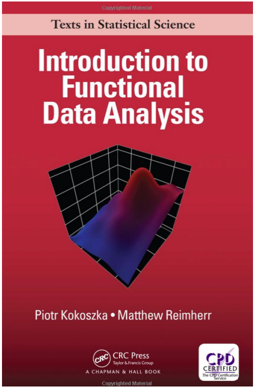


WILEY SERIES IN PROBABILITY AND STATISTICS

Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators

Tailen Hsing • Randall Eubank

WILEY



Copyrighted Material

Texts in Statistical Science

Introduction to Functional Data Analysis

Piotr Kokoszka • Matthew Reimherr

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

CPD CERTIFIED
The CPD Certification Service

Copyrighted Material

What Is Functional Data?

- "*Functional data is multivariate data with an ordering on the dimensions. (Müller, 2006)*"

  $$y_{ij} = x_i(t_{ij}) + \varepsilon_{ij},$$

  with t in a continuum and $x_i(t)$ smooth

- Functional data = the functions $x_i(t), \quad i = 1, \cdots, n$

- Functional Data Analysis (FDA) treats the whole curve as a single entity

Challenges

- Characterization of $X(t)$
  - $\mathbb{E}[X(t)] = \mu(t)$
  - $\mathrm{Var}[X(t)] = \sigma^2(t)$
  - $\mathrm{Cov}\left(X\left(t\right), X\left(t'\right)\right) = \mathrm{C}(t, t')$
- Estimation of $X(t)$ from y_{ij}, noisy and discrete observations
- $n < p = \infty \Rightarrow$ requires regularization with some smoothness assumptions

From Discrete to Functional Data

Two approaches:

- Basis-expansion:

$$x(t) = \sum_{k=1}^{K} \phi_k(t) c_k,$$

  where $\{\phi_i(t)\}$ are pre-chosen basis functions (e.g., B-splines)

- Smoothing penalties:

$$\text{arg max} \left[\sum_{j=1}^{m} \left(y_j - x\left(t_j\right)\right)^2 + \lambda \int \left[\mathcal{L}x(t)\right]^2 \, dt \right],$$

  where λ a “smoothing parameter” and $\mathcal{L}x(t)$ measures the “roughness” of $x(\cdot)$

Functional Principal Component Analysis (FPCA)

- Multivariate PCA, use Eigen-decomposition:

$$\Sigma = \boldsymbol{U}^T \boldsymbol{D} \boldsymbol{U} = \sum_{j=1}^{p} \boldsymbol{d}_j \boldsymbol{u}_j \boldsymbol{u}_j^T$$

- FPCA: use Karhunen-Loève decomposition:

$$\mathrm{C}(t, t') = \sum_{j=1}^{\infty} d_j \xi_j(t) \xi_j(t')$$

- Dimension reduction:

$$x_i(t) \approx \bar{x}(t) + \sum_{j=1}^{K} A_{ij} \xi_j(t),$$

where $A_{ij} = \int \left[x_i(t) - \bar{x}(t) \right] \xi_j(t) \, dt$

Functional Linear Models

- Functional Predictor Regression

$$y_i = \beta_0 + \int \beta(t) z(t)\, dt + \varepsilon_i$$

- Functional Response Regression

$$y_i(t) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) z_{ik} + \varepsilon_i(t)$$

- Function-on-Function Regression

$$y_i(t) = \beta_0(t) + \int \beta(s,t) z_i(s)\, ds + \varepsilon_i(t)$$

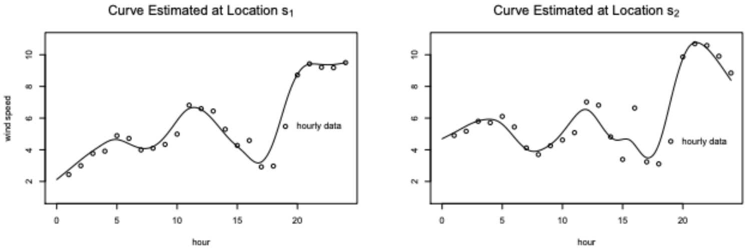
FDA Viewpoint of Space-Time Data

- In Martínez-Hernández & Genton (2020) "*Recent Developments in Complex and Spatially Correlated Functional Data*", they used wind data simulated from the Weather Research Forecasting (WRF) model

- Hourly wind speed at 5-km resolution from 2009-2014 in a 115×115 km^2 region
  $\Rightarrow y(\boldsymbol{s}_i, t_j), \quad i = 1, \cdots, 529, j = 1, \cdots, (365 \times 6 \times 24)$

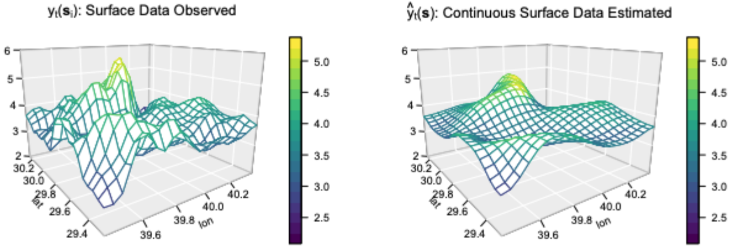
- They examined (very briefly) on the spatial functional data and surface time series viewpoints

Spatial Functional Data Viewpoint



Source: Martínez-Hernández & Genton (2020) Fig. 1

Surface Time Series Viewpoint



$y_t(\mathbf{s}_i)$: Surface Data Observed

$\hat{y}_t(\mathbf{s})$: Continuous Surface Data Estimated

Source: Martínez-Hernández & Genton (2020) Fig. 4

Plans for the Rest of the Semester

- 11/11 Discussion of Parameter estimation for differential equations

- 11/18 Xinyi Li: Image-on-Scalar Regression