Functional data analysis with the `refund` package

Philip T. Reiss
University of Haifa
reiss@stat.haifa.ac.il
http://works.bepress.com/phil_reiss

Psychoco: International Workshop on Psychometric Computing
Dortmund, 27 February 2020

Thanks to . . .

- Co-authors
  - Jeff Goldsmith
  - Fabian Scheipl
  - Lei Huang
  - Julia Wrobel
  - Chongzhi Di
  - Jonathan Gellar
  - Jaroslaw Harezlak
  - Mathew W. McLean
  - Bruce Swihart
  - Luo Xiao
  - Ciprian Crainiceanu
- Daniel Reich for providing diffusion tensor imaging data collected at Johns Hopkins University and the Kennedy Krieger Institute
- Martin Lindquist for providing functional MRI data
- Funding sources including the U.S. National Institutes of Health (National Institute of Mental Health, National Heart, Lung, and Blood Institute, National Institute of Biomedical Imaging and Bioengineering) and the Israel Science Foundation

Outline

Functional data analysis

Splines

refund

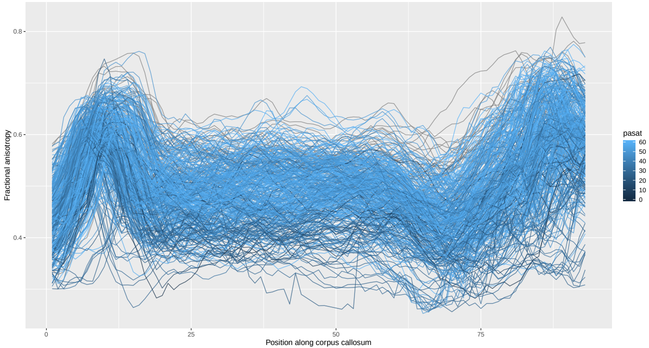
fMRI example

Functional data analysis

- Since the 1990s, a new class of data sets has become common, in which the data for each individual include not just a few measurements, but an entire curve or *function*.
- The term “functional data analysis” (FDA), popularized by Ramsay and Silverman (1997, 2005), refers to methodology for data of this type, which typically extends classical statistical methods (regression, multivariate analysis, etc.)

Example: diffusion tensor imaging (DTI) data

- Each curve represents fractional anisotropy (FA), a measure of white-matter integrity derived by DTI, at 93 locations along the corpus callosum.



- Color denotes PASAT (cognitive function) score—related to FA?
- 142 individuals scanned multiple times—382 observations in total.

CRAN Task View: Functional Data Analysis

Maintainer:	Fabian Scheipl
Contact:	fabian.scheipl at stat.uni-muenchen.de
Version:	2020-02-20
URL:	https://CRAN.R-project.org/view=FunctionalData

Functional data analysis (FDA) deals with data that "provides information about curves, surfaces or anything else varying over a continuum." This task view catalogues available packages in this rapidly developing field.

The R package `refund`* (Reiss et al., 2010; Goldsmith et al., 2019) is a collaborative project implementing methods for

1. functional regression
   - "scalar-on-function" regression: $y \sim x(s)$
   - "function-on-scalar" regression: $y(s) \sim x$
   - "function-on-function" regression: $y(s) \sim x(s)$
2. functional principal component analysis

* short for REgression with FUNctional Data

Why refund?

The original R package `fda` (Ramsay et al., 2009) uses penalized splines to fit functional linear models such as

- the scalar-on-function regression model

$$y_i = \alpha + \int_{\mathcal{S}} x_i(s)\beta(s)ds + \varepsilon_i,$$

  $i = 1, \ldots, n$ (e.g., Ramsay and Silverman, 1997; Marx and Eilers, 1999),
- and the function-on-scalar (varying-coefficient) regression model

$$y_i(s) = \beta_0(s) + x_i\beta_1(s) + \varepsilon_i(s).$$

Limitations:

- restricted to “vanilla” models—without multiple predictors, random effects, extensions to generalized linear models
- smoothing parameter selection is laborious

`refund` lifts these restrictions.

Outline

Functional data analysis

Splines

refund

fMRI example

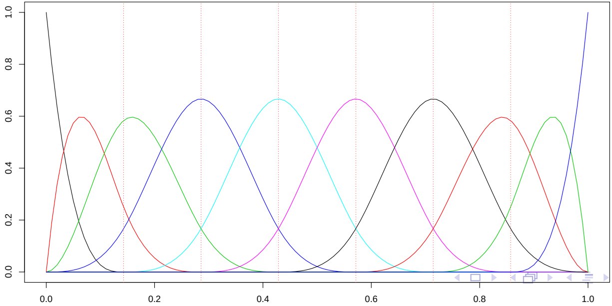
- Penalized splines are a popular way to fit the **nonparametric regression** model

$$y_i = f(x_i) + \varepsilon_i, \qquad E(\varepsilon_i) = 0$$

where f is some smooth function.

- Briefly, the spline approach assumes f to be piecewise polynomial (usually cubic), such that at the “knots” (boundaries) there are a certain number of continuous derivatives (usually 2).
- Specifically, we take f to be a linear combination of B-splines, piecewise polynomial functions with compact support:

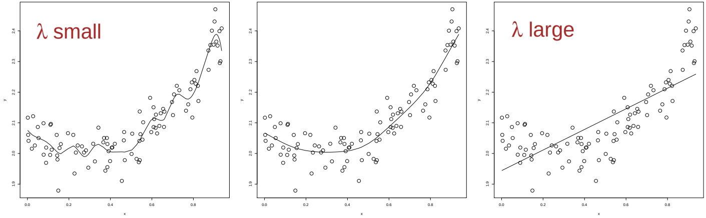
$$f(x) = \boldsymbol{b}(x)^T\boldsymbol{\beta} \text{ where } \boldsymbol{b}(x) = [b_1(x), \ldots, b_K(x)]^T, \boldsymbol{\beta} \in \mathbb{R}^K.$$



- Given a spline basis, we estimate $f(x)$ by penalized least squares, i.e., $\hat{f}(x) = \boldsymbol{b}(x)^T\hat{\boldsymbol{\beta}}$ minimizes

$$\underbrace{\sum_{i=1}^{n}[y_i - f(x_i)]^2}_{\text{sum of squared errors}} + \lambda \underbrace{\int f''(x)^2 dx}_{\text{roughness functional}}$$

  over all functions of the form $f(x) = \boldsymbol{b}(x)^T\boldsymbol{\beta}$.
- Choice of λ is critical:



- Coefficient functions $\beta(s)$ in functional regression are also estimated by (more complicated) penalized least squares.
- `refund` implements fast automatic smoothing parameter selection via the `mgcv` package (Wood, 2011, 2017).

Outline

Functional data analysis

Splines

refund

fMRI example

Regression functions in `refund`

		Predictors	
		Scalar	***Functional***
Responses	***Scalar***		pfr
	Functional	fosr, fosr2s, pffr	pffr

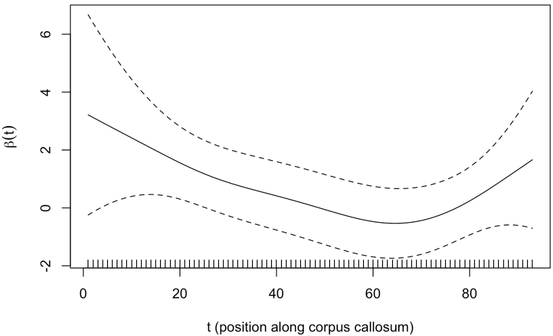
Let's illustrate with the DTI data . . .

Scalar-on-function regression with random subject effects (intercepts):

$$P_{ij} = \alpha_i + \int_{\mathcal{S}} FA_{ij}(s)\beta(s)ds + \varepsilon_{ij},$$

where P is PASAT score and $FA(s)$ is fractional anisotropy curve.

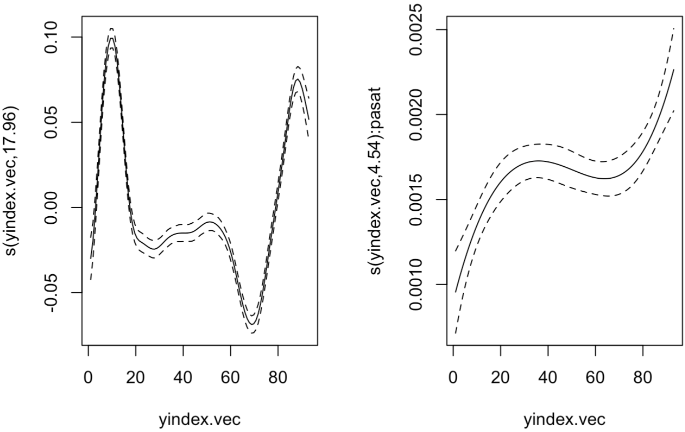
```
library(refund)
data(DTI)
FA.cca <- DTI[complete.cases(DTI$cca),]
FA.cca$ID <- factor(FA.cca$ID)
sofr.fit <- pfr(pasat ~ lf(cca, k=30, argvals=1:93) + re(ID), data=FA.cca)
plot(sofr.fit, select=1, ylab=expression(paste(beta(t))), xlab="t (position along corpus callosum)")
```



A function-on-scalar regression model:

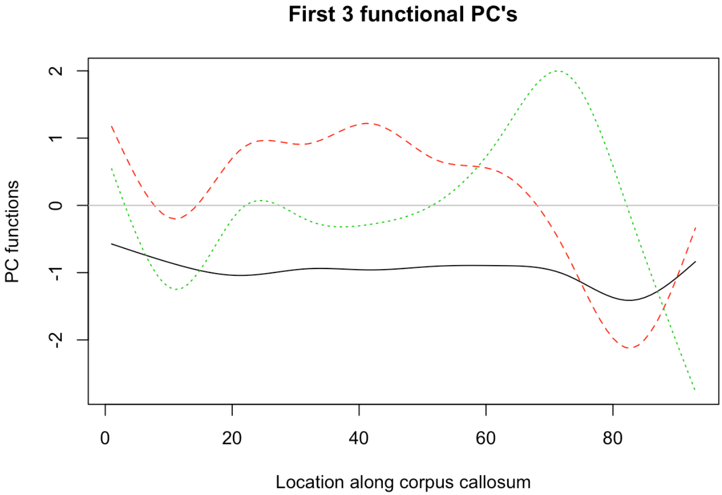
$$FA_{ij}(s) = \beta_0(s) + P_{ij}\beta_1(s) + \varepsilon_{ij}(s).$$

```
fosr.fit <- pffr(cca~pasat, data=FA.cca)
plot(fosr.fit, pages=1, scale=0)
```



Functional PCA:

```
FPC.fit <- fpca.sc(FA.cca$cca, npc=10)
```



The International Journal of Biostatistics

Volume 6, *Issue* 1 2010 *Article* 28

Fast Function-on-Scalar Regression with Penalized Basis Expansions

Philip T. Reiss* Lei Huang†

Maarten Mennes‡

Comput Stat (2015) 30:539–568
DOI 10.1007/s00180-014-0548-4

ORIGINAL PAPER

Penalized function-on-function regression

Andrada E. Ivanescu · Ana-Maria Staicu ·
Fabian Scheipl · Sonja Greven

Biometrics 71, 344–353
June 2015

DOI: 10.1111/biom.122

Generalized Multilevel Function-on-Scalar Regression and Principal Component Analysis

Jeff Goldsmith,[1,*] Vadim Zipunnikov,[2] and Jennifer Schrack[3,4]

Biostatistics (2013), **14**, 3, *pp.* 447–461
doi:10.1093/biostatistics/kxs051
Advance Access publication on January 5, 2013

ELSEVIER
Econometrics and Statistics
journal homepage: www.elsevier.com/locate/ecosta

High-dimensional adaptive function-on-scalar regression

Zhaohu Fan[a], Matthew Reimherr[b,*]
[a] Departments of Industrial Engineering and Statistics, Penn State University, University Park, PA 16802, United States
[b] Department of Statistics, Penn State University, University Park, PA 16802, United States

Longitudinal scalar-on-functions regression with application to tractography data

JAN GERTHEISS*
Department of Animal Sciences, Georg-August-Universität Göttingen, 37075 Gö
jgerthe@uni-goettingen.de
JEFF GOLDSMITH
Department of Biostatistics, Columbia University, New York, NY 1003
CIPRIAN CRAINICEANU
Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21
SONJA GREVEN
Department of Statistics, Ludwig-Maximilians-Universität Munich, 80539 Munich, Germany

Statistical Modelling 2018; **18(3–4)**: 1–19

An introduction to semiparametric function-on-scalar regression

Alexander Bauer[1], Fabian Scheipl[1], Helmut Küchenhoff[1] and Alice-Agnes Gabriel[2]
[1]Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany.
[2]Department of Geophysics, Ludwig-Maximilians-Universität, Munich, Germany.

AStA Advances in Statistical Analysis (2019) 103:411–436
https://doi.org/10.1007/s10182-018-00337-x

ORIGINAL PAPER

A comparison of testing methods in scalar-on-function regression

**Merve Yasemin Tekbudak[1,2] · Marcela Alfaro-Córdoba[3] · Arnab Maity[1]
Ana-Maria Staicu[1]**

Zusammenfassung

In funktionaler Datenanalyse bestehen die Daten aus Funktionen, die auf stetigen Trägern definiert sind. In der Praxis werden funktionale Variablen auf diskreten Gittern beobachtet. Regressionsmodelle sind ein wichtiges Werkzeug, um den Einfluss von Kovariablen auf eine Zielvariable zu modellieren; für funktionale Daten stellen sich besondere Herausforderungen. In dieser Arbeit wird eine generische Modellklasse vorgeschlagen, die Skalar-auf-Funktion, Funktion-auf-Skalar und Funktion-auf-Funktion Regression enthält. Quantilsregression, generalisierte additive Modelle und generalisierte additive

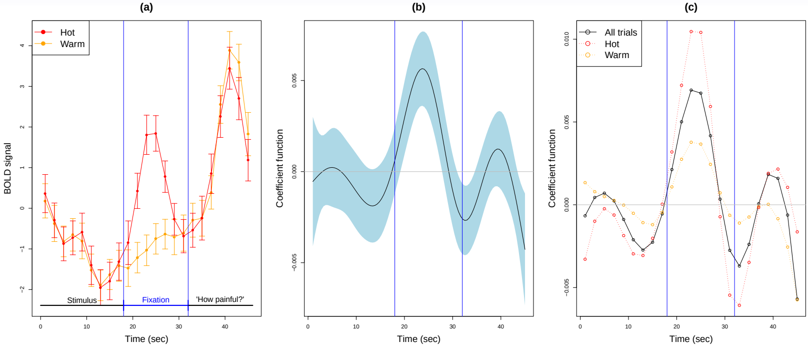
(Brockhaus, 2016)

Outline

- Lindquist (2012) analyzed functional MRI measures of response to pain in 20 volunteers.
- Each volunteer had 39–48 trials consisting of
  - hot (painful) or warm stimulus applied to left forearm (18 sec)
  - a fixation cross on a screen (14 sec)
  - the words "How painful?" appeared on the screen (14 sec)
  - asked to rate the pain intensity on a scale from 100 to 550.
- To study whether BOLD response predicts pain, Reiss et al. (2017) fitted the following scalar-on-function regression model:

$$y_{ij} = \alpha_i + \gamma I_{ij}^{\text{hot}} + \int_{\mathcal{T}} x_{ij}(t)\beta(t)dt + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, J_i,$$

  in which
  - y_{ij} is the log pain score for the ith participant's jth trial;
  - the α_i's are iid normally distributed random intercepts;
  - I_{ij}^{hot} is an indicator for a hot stimulus;
  - $x_{ij}(t)$ is lateral cerebellum BOLD signal over the trial interval $\mathcal{T}$;
  - the ε_{ij}'s are iid normally distributed errors with mean zero.
- γ found to be highly significantly positive; but what about $\beta(t)$?



(a) Mean lateral cerebellum BOLD signal is higher for hot- than for warm-stimulus trials, but only during fixation cross interval.

(b) Coefficient function estimate $\hat{\beta}(t)$, with approximate pointwise 95% confidence intervals.

(c) $\hat{\beta}(t)$ for full data set, versus for only hot or only warm trials.

Interpretation of the peak in $\hat{\beta}(t)$: A "brain signature" for pain?

- A possible explanation is collinearity, or confounding, between $\gamma I_{ij}^{\text{hot}}$ (painful heat) and $\int_{\mathcal{T}} x_{ij}(t)\beta(t)dt$ (BOLD signal effect).
- But since $\hat{\beta}(t)$ looks similar when restrict to each of two temperature conditions [subfigure (c) on previous slide], it may be that brain activity partially mediates the painful effect of the hot stimulus.

More to explore . . .

- The `refund.shiny` package (Wrobel et al., 2016) offers interactive graphics for various analyses with functional data.

- Chapter 13 of Mair (2018) discusses function-on-scalar regression with `refund` applied to psychometric data.
- The monograph of Kokoszka and Reimherr (2017) on functional data analysis includes many `refund` examples.

Thank you!

Photo: Berthold Werner

References

Brockhaus, S. (2016). *Boosting functional regression models*. Ph. D. thesis, Ludwig-Maximilians-Universität München.

Goldsmith, J., F. Scheipl, L. Huang, J. Wrobel, C. Di, J. Gellar, J. Harezlak, M. W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, and P. T. Reiss (2019). *refund: Regression with Functional Data*. R package version 0.1-21.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. CRC Press.

Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association 107*, 1297–1309.

Mair, P. (2018). *Modern Psychometrics with R*. Springer.

Marx, B. D. and P. H. C. Eilers (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics 41*(1), 1–13.

Ramsay, J. O., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. New York: Springer.

Ramsay, J. O. and B. W. Silverman (1997). *Functional Data Analysis*. New York: Springer.

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

Reiss, P. T., J. Goldsmith, H. L. Shang, and R. T. Ogden (2017). Methods for scalar-on-function regression. *International Statistical Review 85*(2), 228–249.

Reiss, P. T., L. Huang, and M. Mennes (2010). Fast function-on-scalar regression with penalized basis expansions. *International Journal of Biostatistics 6*(1, article 28).

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B 73*(1), 3–36.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Boca Raton, Florida: CRC Press.

Wrobel, J., S. Y. Park, A. M. Staicu, and J. Goldsmith (2016). Interactive graphics for functional data analyses. *Stat 5*(1), 108–118.