Compositional functional regression with isotemporal substitution and its application to time-use behavioural data

P. Jašková[1], K. Hron[1], A. Gába[2]

[1]Faculty of Science, Palacký University Olomouc, Czech Republic
[2]Faculty of Physical Culture, Palacký University Olomouc, Czech Republic

16. 9. 2021

Data as functions

- functional data analysis (FDA) processes information about data, which have functional character (Ramsay and Silverman, 2005)
- functional data are entities that can be described through a function (curve)
- functional dataset consists of a sample of functional observations
- in the statistical analysis of functional data, often the observations can be characterised as probability density functions (PDFs)

Probability density functions

- standard methods of FDA (Ramsay and Silverman, 2005) are not suitable for statistical processing of density functions due to their relative nature (> 0 and $\int_I = 1$)
- compositional data – multivariate observations carrying relative information
- any other representative within the class of proportional PDFs would carry the same set of relative information (scale invariance)
- PDFs can be represented with a unit integral constraint without loss of information
- probability density functions = functional compositional data (functional data carrying only relative information)
- geometrically represented using Bayes spaces

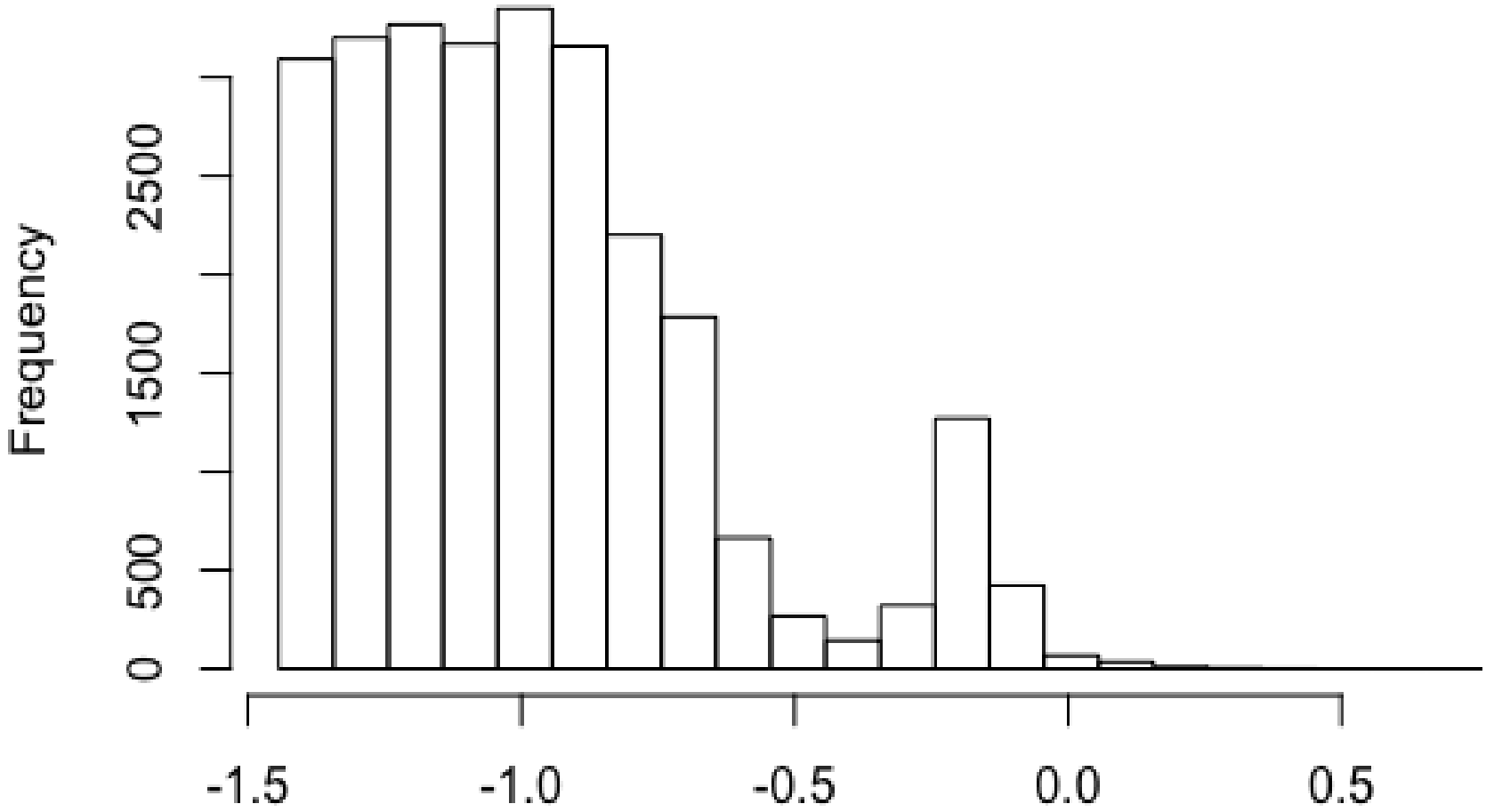
Bayes spaces

- $\mathcal{B}^2(I)$ . . . . . the Bayes space of nonnegative functional compositions on a compact subset I of $\mathbb{R}$ (usually an interval)
- operations of perturbation and power transformation ($f, g \in \mathcal{B}^2(I)$, $\alpha \in \mathbb{R}$):

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)\,\mathrm{d}s}, \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha\,\mathrm{d}s}, \quad t \in I$$

Preprocessing of PDFs

- First step:



Representation of functional data

Model:

$$f_i = x(t_i) + \epsilon_i, i = 1, \ldots, N,$$

- ϵ_i . . . functional error considered in the model
- further assumption: smoothness of the function (existence of continuous derivatives of a certain order)

Observed functional data are N pairs (f_i, t_i) , in our case they correspond to representatives of histogram classes

- possible ways to determine functional form:
  - **interpolation**: a process of obtaining functions from the measured data, if our observations do not contain an error
  - **approximation**: observations with error (used for smoothing)

Basis functions

System of basis functions is a set of known functions that are linearly independent and allows to approximate well any function as a linear combination of K these functions $\{\varphi_1, \ldots, \varphi_K\}^T$

We will express a function $x(t)$ by the linear expansion

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

where φ_k are known basis functions and $(c_1, \ldots, c_K)^T$ is a vector of unknown coefficients

Well-known basis expansions

- well-known basis expansions:
  - *Fourier series*: used typically for periodic data
  - *B-spline*: for non-periodic functional data
- B-spline basis functions also allow to capture complex behavior in large datasets

B-spline representation

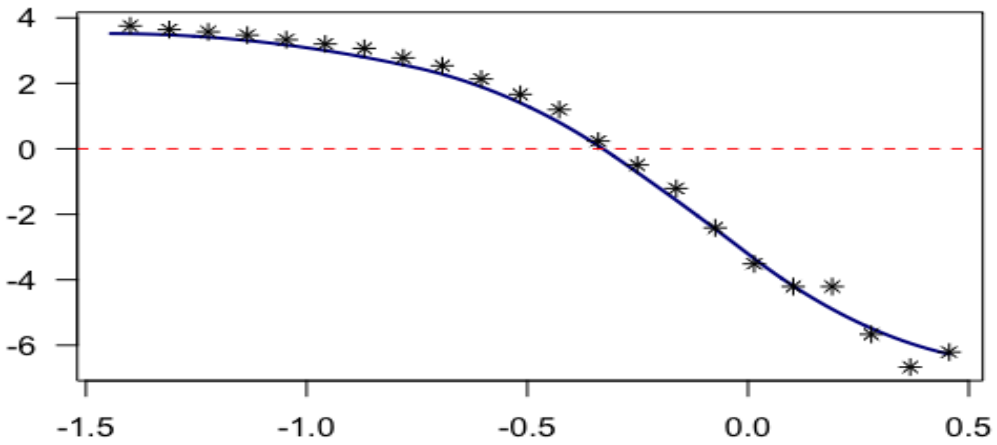
- A B-spline is a polynomial function of low degree comprising several parts with the individual parts following each other sufficiently smoothly
- every polynomial spline $s_k(t)$ can be uniquely expressed as

$$s_k(t) = \sum_{i=-k}^{g} b_i B_i^{k+1}(t),$$

  - $b = (b_{-k}, \ldots, b_g)^T$, vector of B-spline coefficients of spline $s_k(t)$
  - $B_i^{k+1}(t)$, B-spline of degree k

Smoothing spline

- method for data approximation
- compromise between interpolation by splines and approximation of data in the sense of the least squares method
- the functional form is smoother than the actual observations



Probability density functions

- problem: standard FDA methods are not suitable for probability densities
  - not capturing geometric properties of PDFs
- solution: using transformation from Bayes space to L^2 space $\rightarrow$ clr transformation
  - centered logratio transformation:

$$\mathrm{clr}(f)(t) := \ln f(t) - \frac{1}{\eta}\int_I \ln f(t)dt$$

  - zero integral constraint which needs to be taken into account also for spline approximation

Probability density functions

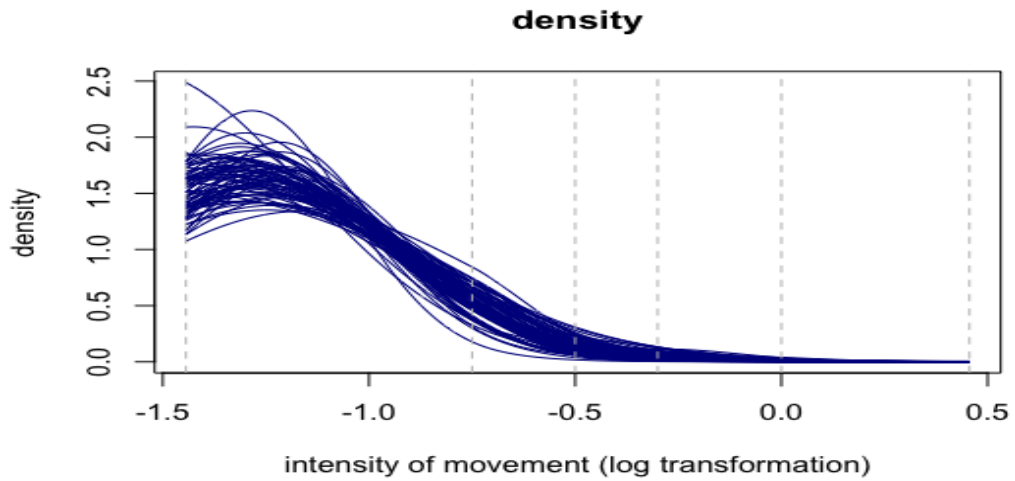
- **Aim**: find a spline with zero integral in $I = [a, b]$:

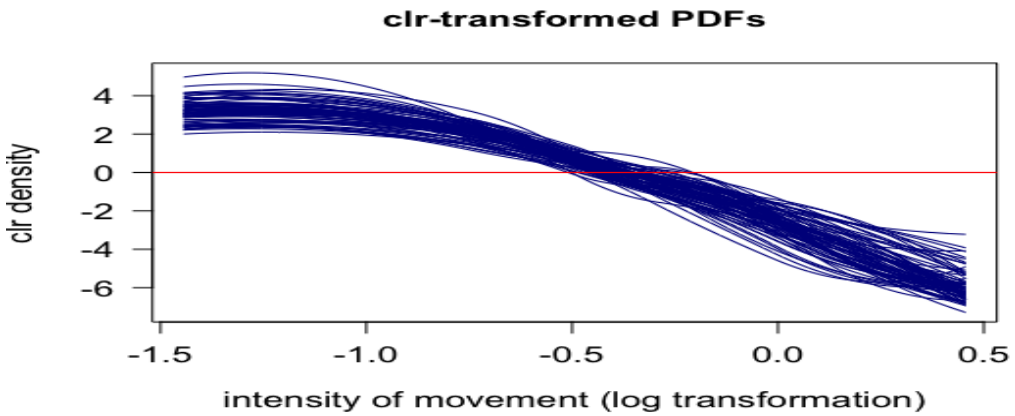
$$\int_a^b s_k(x)dx = 0$$

  $\Longrightarrow$ special type of function usually with zero integral (Machalová and Talská, 2021)
- Application: time-use relative data, compositional data with parts corresponding to physival activity (PA) categories. This can be extended to the continuous case as probability density functions (PDFs).

Illustrative data

- 74 school-aged children (girls between 14 and 17)
- daily activity usually expressed as 5 PA categories (sleep, sedentary behaviour, light PA, moderate PA, vigorous PA), but in fact this results from post-discretization of data which have functional character (PDFs)





Illustarive data

- Question of interest: we are interested in relationship between the PDF expressing movement intensity and a health outcome
- PA continuum expressed as probability density functions (PDFs)
  - relationship $\rightarrow$ regression of health outcome on movement intensity
  - predictor: PDFs
  - real response: percentage body fat mass

  $\rightarrow$ **use of standard functional regression model with real-valued response in the clr space by honouring the zero integral constraint**

SFPCA – Simplicial functional principal component analysis

- Problem: the total number of basis functions will exceed the number of observations (N)
- aim of SFPCA is to reduce the dimensionality of dataset
- replace the original set of observations (functions) with fewer latent functions to explain as much of the variability in the data as possible
- they are called functional principal components
- 2 components explain 94.18% of the total variability in case study

Functional regression model

- functional linear regression model for ith observation y_i with functional predictor f_i is expressed as

$$y_i = \beta_0 + \int_I \beta_1(t)\, f_i(t)dt + \epsilon_i, \quad i = 1, \ldots, N, \quad t \in I$$

- $\int_I \beta_1(t)\, f_i(t)dt = \langle \beta_1(t), f_i(t) \rangle_2$, inner product
  - $\beta_0 \in \mathbb{R}$, scalar intercept
  - $\beta_1(t) \in \mathcal{B}^2(I)$, functional regression parameter
- similar to standard regression model: estimators $\hat{\beta}_0$,$\hat{\beta}_1$ minimize

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \int_I \beta_1(t)\, f_i(t)dt)^2$$

Functional regression model

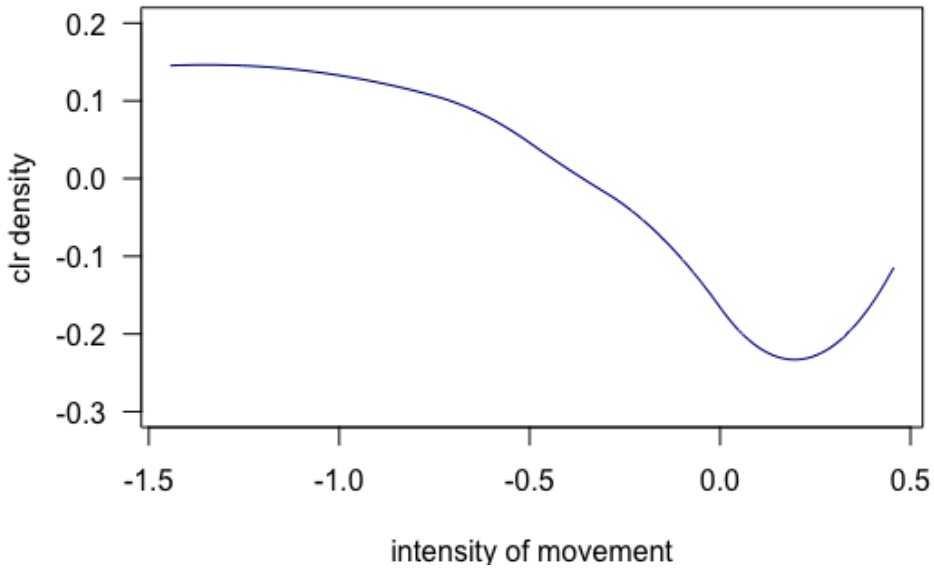
- the regression model can be formulated in the clr space as (Talska and Hron, 2021)

$$y_i = \beta_0 + \int_I \mathrm{clr}(\beta_1)(t)\mathrm{clr}(f_i)(t)\,\mathrm{d}t + \epsilon_i$$

- $\int_I \mathrm{clr}(\beta_1)(t)clr(f_i)(t)\,\mathrm{d}t = \langle clr(\beta_1)(t), \mathrm{clr}(f_i)(t)\rangle_2$
- similar estimation of parameters: estimators $\hat{\beta}_0$,$\hat{\beta}_1$ minimize

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \langle \mathrm{clr}(\beta_1)(t), \mathrm{clr}(f_i)(t)\rangle_2)^2$$

- estimate $\hat{\beta}_1$ is a function, its interpretation is preferred in the clr space



- positive functional values of the regression parameter contribute to the growth of the values of the percentage of fat
- negative values have the opposite effect

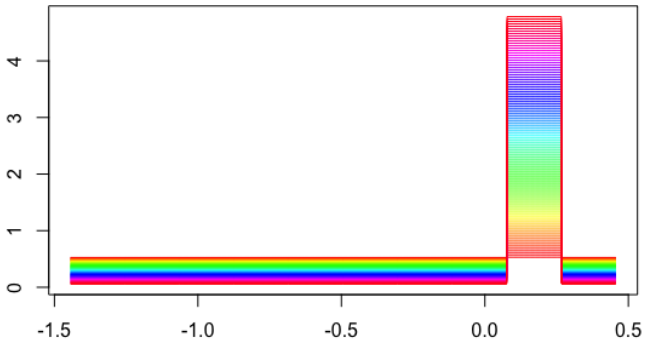
$\rightarrow$ with higher daily activity the percentage of fat decreases (with some artifact effect in higher intensities resulting from approximation of sparse histogram data)

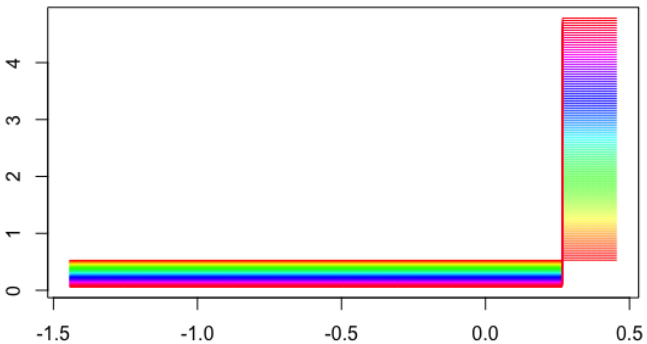
Isotemporal substitution

- **idea**: describe effect of certain subintervals of the accelerometry PDF $\rightarrow$ isotemporal substitution
- model:

$$y = \beta_0 + \left\langle \beta_1^*, \bar{f} \right\rangle_2$$

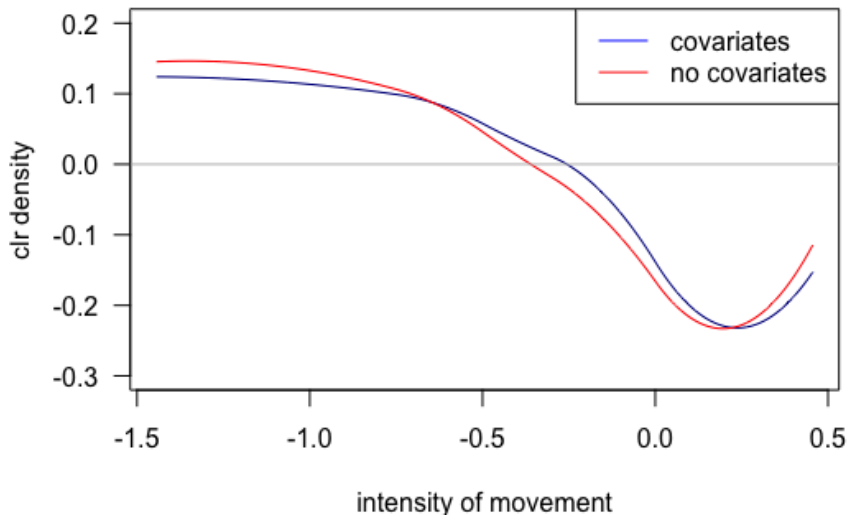
  - $\beta_1^* = \beta_1 + g$
  - $\bar{f}$. . . geometric mean (centre) of PDFs
  - g . . . "weighting" PDF



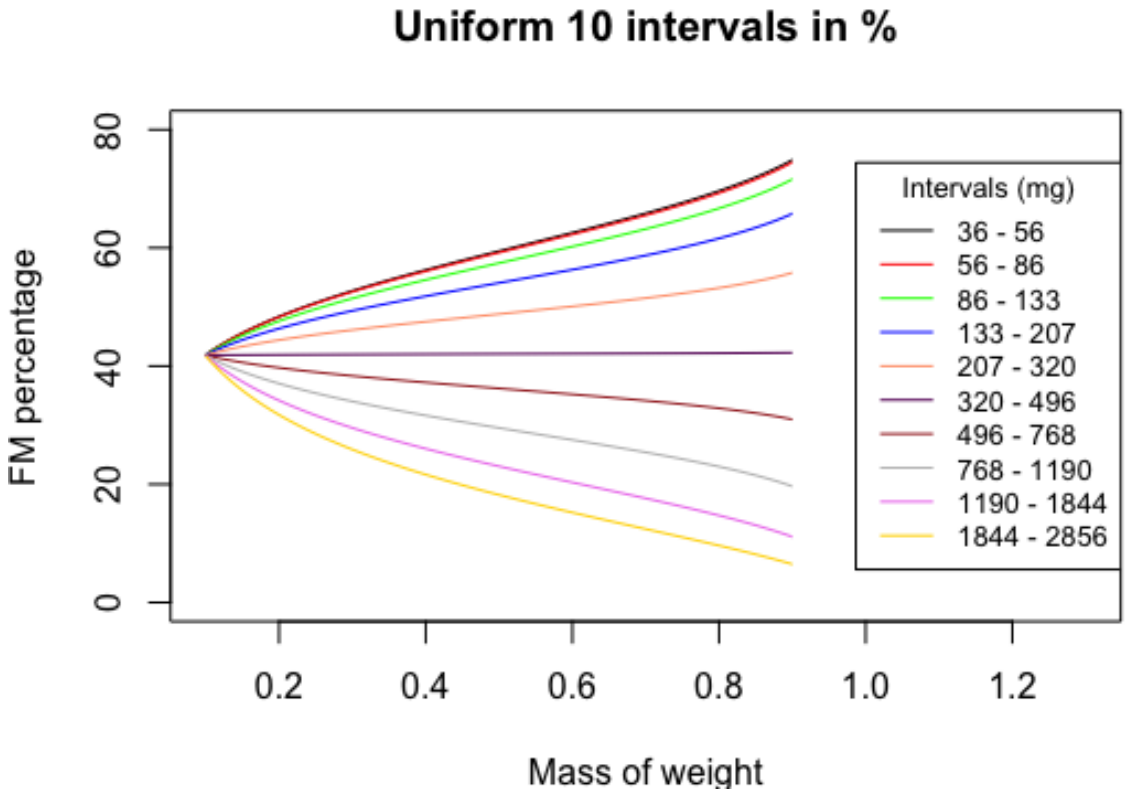


Model with covariates

- problem: hard to distinguish between sleep and sendentary behaviour from the accelerometer
- make three-part composition:
  - first part: take values higher than 36 mg
  - sleep and SB are considered together in the first step of SBP - this leads to two orthonormal coordinates which were added as covariates to the regression model



Results



- curves express percentages of body fat
- result: more time spent in higher intensity PA is associated with lower percentage of fat mass

Thank you for your attention.

References

- K. Hron, A. Menafoglio, M. Templ, K. Hrůzová, P. Filzmoser: Simplicial principal component analysis for density functions in Bayes spaces. Computational Statistics and Data Analysis, 94, 330-350, 2016.
- J. Machalová, R. Talská, K. Hron, A. Gába: Compositional splines for representation of density functions. Computational Statistics 36, 1031–1064, 2021.
- J.O. Ramsay, B.W. Silverman: Functional Data Analysis. Second edition. Springer, New York, 2005
- K.G. van den Boogaart, J.J. Egozcue, V. Pawlowsky-Glahn: Bayes Hilbert spaces. Australian and New Zealand Journal of Statistics 56, 171-194, 2014.
- R. Talská, A. Menafoglio, J. Machalová, K. Hron, E. Fišerová: Compositional regression with functional response. Computational Statistics and Data Analysis 123, 66–85 ,2018.
- R. Talská, K. Hron, T. Matys Grygar: Compositional scalar-on-function regression with application to sediment particle size distributions. Mathematical Geosciences, DOI: 10.1007/s11004-021-09941-1, 2021.