Compositional regression with functional response

R. Talská[a], A. Menafoglio[b*], J. Machalová[a], K. Hron[a], E. Fišerová[a]

[a]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, CZ-77146 Olomouc, Czech Republic,
[b]MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Abstract

The problem of performing functional linear regression when the response variable is represented as a probability density function (PDF) is addressed. PDFs are interpreted as functional compositions, which are objects carrying primarily relative information. In this context, the unit integral constraint allows to single out one of the possible representations of a class of equivalent measures. On these bases, a function-on-scalar regression model with distributional response is proposed, by relying on the theory of Bayes Hilbert spaces. The geometry of Bayes spaces allows capturing all the key inherent features of distributional data (e.g., scale invariance, relative scale). A B-spline basis expansion combined with a functional version of the centred log-ratio transformation is utilized for actual computations. For this purpose, a new key result is proved to characterize B-spline representations in Bayes spaces. The potential of the methodological developments is shown on simulated data and a real case study, dealing with metabolomics data. A bootstrap-based study is performed for the uncertainty quantification of the obtained estimates.

Keywords: Bayes spaces, regression analysis, density functions, B-spline representation

1. Introduction

Distributional data in their discrete form frequently occur in many real-world surveys. For instance, frequencies of occurrence of observations from a continuous random variable – aggregated according to a given partition of the domain of observation – are typically represented by a histogram, which in turn approximates an underlying (continuous) probability density function (PDF).

[1]*Corresponding author. Alessandra Menafoglio, MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy. Email: alessandra.menafoglio@polimi.it.

[2]The codes that implement the proposed regression methods are available on-line as supplementary material of the present article.

In general, PDFs are Borel measurable functions that are constrained to be non-negative and to integrate to unity. One may think at the unit-integral constraint as a way to single out a proper representation of the underlying measure rather than an inherent feature of PDFs themselves. Indeed, when changing the value to which the PDF integrates, to a general positive constant c (i.e., the measure of the whole), the *relative* information carried by PDFs is preserved, this property being called *scale invariance* of PDFs. Here, *relative information* is to be interpreted in terms of the contributions of Borel sets of real line to the overall measure of the support of the corresponding random variable (Hron et al., 2016). Due to the peculiar features of PDFs (e.g., the aforementioned scale invariance and additional properties such as the so-called *relative scale*) the standard L^2 space of square integrable functions appears to be inappropriate for their representation. For instance, the sum of two PDFs according to the geometrical structure of the L^2 space leads to a function that is not a PDF anymore. Even more interestingly, multiplication of a PDF by a real constant yields a scaled PDF, which carries the same relative information as the original PDF according to scale invariance. The relative nature of PDFs indicates that *ratios* between values rather than absolute values represent the relevant source of information. Accordingly, instead of absolute differences, ratios between them should be considered to measure distances and dissimilarities.

In this context, Bayes (Hilbert) spaces provide a well-defined geometrical framework to represent PDFs (van den Boogaart et al., 2010, 2014; Egozcue et al., 2006). The idea motivating the introduction of Bayes spaces was to generalize the well-known Aitchison geometry for finite-dimensional compositional data (i.e., positive observations carrying exclusively relative information, Aitchison (1986); Pawlowsky-Glahn et al. (2015)) to the infinite-dimensional setting. In fact, any PDF can be seen as a composition with infinitely many parts.

Although the general problem of functional regression has been extensively studied in the literature on functional data analysis (FDA, e.g., Ramsay and Silverman, 2005), to the best of the authors' knowledge none of the available works propose a concise methodology for regression analysis in the presence of a distributional response. In this context, this work aims to develop a general theoretical and computational setting allowing for the estimation and uncertainty assessment in linear models with a distributional response. This is relevant from both the methodological and the application-oriented viewpoints. Indeed, having at one's disposal a statistical methodology for the regression of PDF data would enable to assess the entire distribution of the response variable, rather than few statistical moments, such as the mean and the variance. Besides, it would constitute a valuable alternative to quantile regression, with the significant advantage of (a) assessing all the distribution's quantiles jointly and (b) guaranteeing that the ordering among quantiles is preserved by the estimation procedure.

The key point of the proposed approach is to consider PDFs as elements of a Bayes space, and accordingly work with the geometry of the latter space. The centred log-ratio (clr) transformation – that allows representing the PDFs through zero-integral elements of L^2 – is then used to ease computations while

using the Bayes space geometry (van den Boogaart et al., 2014; Hron et al., 2016; Menafoglio et al., 2014, 2016a,b). A B-spline representation of clr-transformed data (Machalová et al. (2016)) is employed to express discretely observed PDFs as smooth functions. On these bases, effective computational procedures are proposed to perform the estimations and assess their uncertainty. The potential of the proposed method shall be demonstrated through a real case study dealing with metabolite concentrations. Further, a simulation study will be introduced to assess the sensitivity of the methodology to the parameters associated with the B-spline representation (e.g., number of knots).

The remaining part of the work is organized as follows. Section 2 recalls the basic notion of Bayes spaces as mathematical spaces for PDF data. The function-on-scalar regression model is briefly recalled in Section 3 for data in L^2. A function-on-scalar model for distributional responses in Bayes spaces is discussed in Section 4. Section 5 proposes a novel computational setting – based on a B-spline representation for PDFs in Bayes spaces – which can be employed for actual computations of the proposed estimators, while Section 6 relates our findings with previous works on compositional regression for multivariate data. Section 7 tests the performances of the method through an extensive simulation study. Section 8 illustrates the application of the methodological developments to real data on metabolites concentrations, and Section 9 finally concludes the work.

2. Probability densities as elements of Bayes spaces

As for finite-dimensional compositional data, a proper choice of the sample space for PDFs is essential. Indeed, as shown in Delicado (2011) and Hron et al. (2016), analysing PDFs within the usual L^2 space may lead to meaningless results. Instead, the peculiarities of densities can be captured through Bayes spaces, which rely upon an appropriate Hilbert space structure to deal with the data constraints.

We consider two positive functions f and g with the same support to be equivalent if $f = c \cdot g$, for a positive constant c. Recalling the scale invariance of PDFs, this implies that densities (not necessarily unit-integral densities, i.e., PDFs) within an equivalence class provide the same relative information, or, equivalently, which contributions of Borel sets to the whole mass measure do not change. For a density f, we denote by $\mathcal{C}(f)$ the unit-integral representative within its equivalence class, also named *closure*. The Bayes space $\mathcal{B}^2(I)$ consists of (equivalence classes of) densities f on a domain I for which the logarithm is square-integrable. Although the theory of van den Boogaart et al. (2014) is general and allows dealing with unbounded supports I, its construction for non-compact supports relies on reference measures different from the Lebesgue one. The latter general case raises foundational issues – both methodological and practical – which are still open. For the purpose of this work, the focus is here on the case of a compact support $I = [a, b] \subset \mathbf{R}$, which was demonstrated to be of broad applicability by several authors (Delicado, 2011; Hron et al., 2016; Menafoglio et al., 2014, 2016a,b).

In $\mathcal{B}^2(I)$, the counterparts of sum and multiplication by a scalar are called *perturbation* and *powering*, and are defined, for $f, g \in \mathcal{B}^2(I)$ and $c \in \mathbf{R}$, as

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_a^b f(s)g(s)ds} = \mathcal{C}(fg)(t); \qquad (c \odot f)(t) = \frac{f^c(t)}{\int_a^b f^c(s)ds} = \mathcal{C}(f^c)(t),$$

where $t \in I = [a, b]$. Note that $e(t) = \frac{1}{b-a}$ (uniform density on $[a, b]$) is the neutral element of perturbation. The Bayes inner product is finally defined as

$$\langle f, g \rangle_{\mathcal{B}} = \frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{f(t)}{f(s)} \cdot \ln \frac{g(t)}{g(s)} dt\, ds,$$

where η stands for the length of interval I, i.e., $\eta = b - a$. The corresponding norm and distance are

$$\|f\|_{\mathcal{B}} = \sqrt{\langle f, f \rangle_{\mathcal{B}}}; \quad d_{\mathcal{B}}(f, g) = \|f \ominus g\|_{\mathcal{B}},$$

respectively, where $\ominus$ stands for *perturbation-subtraction* of f by g, $(f \ominus g)(t) = [f \oplus (-1) \odot g](t)$, for t in I. Egozcue et al. (2006) and van den Boogaart et al. (2014) showed that the Bayes space $\left(\mathcal{B}^2(I), \oplus, \odot, \langle \cdot, \cdot \rangle_{\mathcal{B}}\right)$ forms a separable Hilbert space.

For a given compact support I there exists an isometric isomorphism between the Bayes space $\mathcal{B}^2(I)$ and the space $L^2(I)$ of square integrable real functions on I. An instance of such isometric isomorphism is called *centred log-ratio (clr) transformation*, defined, for a PDF $f \in \mathcal{B}^2(I)$,

$$f_c(t) = \operatorname{clr}[f](t) = \ln f(t) - \frac{1}{\eta} \int_I \ln f(s)\, ds, \quad t \in I. \tag{1}$$

The clr representation of a PDF is featured by a zero-integral constraint on I, i.e., $\int_I f_c(t)dt = 0$. When analyzing clr transforms of densities, the latter integral constraint may give rise to computational issues and thus needs to be properly accounted for. Given a clr transform $f_c \in L^2(I)$, the corresponding density $f \in \mathcal{B}^2(I)$ can be obtained through the inverse transformation, $f(t) = \operatorname{clr}^{-1}[f_c](t) = \mathcal{C}(\exp[f_c])(t)$, $t \in I$, $\mathcal{C}$ denoting the closure. Finally, we point out that the following important properties of the isometric isomorphism (1) hold

$$\operatorname{clr}(f \oplus g)(t) = f_c(t) + g_c(t), \quad \operatorname{clr}(c \odot f)(t) = c \cdot f_c(t), \quad \langle f, g \rangle_{\mathcal{B}} = \langle f_c, g_c \rangle_2,$$

where $\langle \cdot, \cdot \rangle_2$, $\|\cdot\|_2$ and $d_2(\cdot, \cdot)$ denote the inner product, norm and distance in $L^2(I)$ respectively. Intuitively, the clr transformation translates operations and metrics of the Bayes space into the usual operations and metrics of the L^2 space.

3. Functional regression model for unconstrained data in L^2

A large body of literature has been developed on both theoretical and applied issues related to functional linear models (Faraway, 1997; Ramsay and

Silverman, 2005; Shena and Xub, 2007). We here review the key notions on function-on-scalar regression that are deemed useful for our developments, by following Ramsay and Silverman (2005, Chapter 13), to which readers are referred for further details.

A function-on-scalar regression model relates a functional response $y(t)$ with independent scalar covariates x_j for $j = 0, \ldots, r$, the first regressor x_0 indicating the intercept, $x_0 = 1$. Consider an N-dimensional vector of functional observations $\mathbf{y}(t)$ in $L^2(I)$, a design matrix $\mathbf{X}$ of dimension $N \times p$ (the first column being $(1, ..., 1)' \in \mathbb{R}^N$ if the intercept is included) and a p-dimensional vector of unknown functional regression parameters $\boldsymbol{\beta}(t)$ in $L^2(I)$. Call $\boldsymbol{\varepsilon}(t)$ an N-dimensional vector of i.i.d. (functional) random errors with zero-mean in $L^2(I)$. The functional linear model for the i-th observation y_i, $i = 1, ..., N$, associated with the regressors x_{ij}, $j = 0, ..., r$, is expressed as

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{r} x_{ij}\beta_j(t) + \varepsilon_i(t), \quad i = 1, \ldots, N, \tag{2}$$

or, in matrix notation, $\mathbf{y}(t) = \mathbf{X}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t)$, where $p = r + 1$ and $x_{i0} = 1$. The estimators $\widehat{\beta}_j$, $j = 0, ..., r$, of the coefficients β_j, $j = 0, ..., r$, can be obtained by minimizing the least square fitting criterion,

$$\mathrm{SSE}(\boldsymbol{\beta}) = \int_I [\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)]' [\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)] \, dt. \tag{3}$$

The smoothness of the resulting estimations may be controlled by adding a differential penalization to the SSE criterion, i.e.,

$$\mathrm{PENSSE}(\boldsymbol{\beta}) = \int_I [\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)]' [\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)] \, dt + \lambda \int_I [L\boldsymbol{\beta}(s)]' [L\boldsymbol{\beta}(s)] \, ds, \tag{4}$$

with L a linear differential operator and λ a smoothing parameter.

Several computational methods have been proposed in the literature to minimize (3) or (4). In Ramsay and Silverman (2005) methods relying upon basis expansions of the functional observations $y_i(t)$, $i = 1, \ldots, N$, and regressors $\beta_j(t)$, $j = 0, \ldots, r$, are broadly discussed. The latter are briefly recalled in Appendix A.

4. Functional regression when the response is a density

In this section, a functional regression model in $\mathcal{B}^2(I)$ is introduced as a counterpart of the model (2). We assume the dependent variable $y(t)$ to be an element of $\mathcal{B}^2(I)$ and consider scalar covariates x_j, $j = 0, \ldots, r$. Each observation of the distributional response $y_i(t)$, $i = 1, \ldots, N$, is thus associated with a vector of p covariates, $x_{i0}, \ldots, x_{ir}$, with $x_{i0} = 1$ for $i = 1, ..., N$. We consider a functional linear model in $\mathcal{B}^2(I)$ of the form

$$y_i(t) = \beta_0(t) \oplus \bigoplus_{j=1}^{r} [x_{ij} \odot \beta_j](t) \oplus \varepsilon_i(t) \tag{5}$$

where ε_i denotes a zero-mean functional error (or residual) in $\mathcal{B}^2(I)$, $i = 1, \ldots, N$, and the unknown functions β_j, $j = 0, ..., r$, belong to $\mathcal{B}^2(I)$ as well. To estimate the coefficients $\beta_j(t), j = 0, \ldots, r$, we minimize the functional sum of square-norms of the error in $\mathcal{B}^2(I)$

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\varepsilon_i\|_{\mathcal{B}}^2 = \sum_{i=1}^{N} \left\| \bigoplus_{j=0}^{r} [x_{ij} \odot \beta_j] \ominus y_i \right\|_{\mathcal{B}}^2. \tag{6}$$

Note that (6) is the counterpart of SSE (3) in the Bayes Hilbert space; in fact, it also represents the analogue of compositional SSE (Egozcue et al., 2012) in infinite dimensions. Applying the clr transformation (1) to both sides of the model (5) yields

$$\mathrm{clr}(y_i)(t) = \mathrm{clr}(\beta_0)(t) + \sum_{j=1}^{r} \left[x_{ij} \cdot \mathrm{clr}(\beta_j)\right](t) + \mathrm{clr}(\varepsilon_i)(t), \quad i = 1, \ldots, N, \tag{7}$$

that enables one to reformulate the objective SSE (6) equivalently in the L^2 sense as

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\mathrm{clr}(\varepsilon_i)\|_2^2 = \sum_{i=1}^{N} \left\| \sum_{j=0}^{r} [x_{ij} \cdot \mathrm{clr}(\beta_j)] - \mathrm{clr}(y_i) \right\|_2^2. \tag{8}$$

In this work, the focus is on SSE, since one may control the smoothness of the estimated functions for $\mathrm{clr}(\beta_j(t))$ through the smoothness of the B-spline representation of the response, as shall be discussed in Section 6. Note that alternatively one could develop PENSSE, by closely follow the arguments here presented.

Both the clr of observed functions $\mathrm{clr}(y_i)(t)$, $i = 1, \ldots, N$, and of regression coefficients $\mathrm{clr}(\beta_j)(t)$, $j = 0, \ldots, r$, in (8) need to follow the zero-integral constraint, i.e.,

$$\int_I \mathrm{clr}(y_i(t))\, dt = 0; \quad \int_I \mathrm{clr}(\beta_j(t))\, dt = 0,\ j = 0, ..., r. \tag{9}$$

In the following, we will use a basis representation for both $\mathrm{clr}(y_i(t))$, $i = 1, \ldots, N$, and $\mathrm{clr}(\beta_j)(t)$, $j = 0, \ldots, r$, as detailed in Section 5.1. Let $\{\varphi_k, k = 1, ..., K\}$ be a given basis system and let us express $\mathrm{clr}(y_i)(t)$, $i = 1, \ldots, N$, and $\mathrm{clr}(\beta_j)(t)$, $j = 0, \ldots, r$, on such basis as

$$\mathrm{clr}(y_i(t)) = \sum_{k=1}^{K} c_{ik}\varphi_k(t); \quad \mathrm{clr}(\beta_j(t)) = \sum_{k=1}^{K} b_{jk}\varphi_k(t), \tag{10}$$

or, in matrix notation, $\mathrm{clr}(y_i(t)) = \mathbf{c}_i'\boldsymbol{\varphi}(t)$ and $\mathrm{clr}(\beta_j(t)) = \mathbf{b}_j'\boldsymbol{\varphi}(t)$. Then, the zero-integral constraints in (9) read

$$\int_I \mathrm{clr}(y_i(t))dt = \int_I \sum_{k=1}^{K} c_{ik}\varphi_k(t)dt = 0; \quad \int_I \mathrm{clr}(\beta_j(t))dt = \int_I \sum_{k=1}^{K} b_{jk}\varphi_k(t)dt = 0. \tag{11}$$

These constraints need to be carefully taken into account when estimating the linear model (5), as they may turn in linear constraints on the coefficients $\{c_{ik}\}, \{b_{jk}\}$ and consequently on model singularities. We discuss this point and its implications in the next Sections, in the light of the key result proved in Section 5.1.

5. Function-on-scalar regression for densities represented via B-splines in Bayes spaces

In this section, we briefly recall the basic notions on smoothing B-splines for density functions following Machalová et al. (2016), and show that the zero-integral constraint on clr induces a constraint on B-spline coefficients which is characterizing of this class of B-splines. This will be used to propose a method to overcome the problem of singularity in the regression model.

5.1. The B-spline representation for density functions in Bayes spaces

In most practical situations, the PDFs under study are discretely sampled in terms of histogram data. That is, for each of the densities $y_i(t)$, $t \in [a, b]$, $i = 1, \ldots, N$, one usually observes a positive real vector $\mathbf{W}_i = (W_{i1}, \ldots, W_{iD})'$, whose components correspond to the (absolute or relative) frequencies of the classes in which the interval I is partitioned; possible count zeros can be effectively replaced by using methods from Martín-Fernández et al. (2015). Note that vectors $\mathbf{W}_i$, $i = 1, ..., N$, are constrained similarly as the PDFs $y_i, i = 1, ..., N$. They can be interpreted as compositional data, and analysed by using similar ideas as in Bayes spaces (Pawlowsky-Glahn et al., 2015). In order to express these vectors in a standard Euclidean space, one may employ the discrete version of the clr transformation (1) (e.g., Pawlowsky-Glahn et al., 2015). Denote by $\underline{\mathbf{Z}} = (Z_{ij})$ the matrix of clr-transformed raw data. To estimate the underlying continuous density from raw data we here consider the smoothing splines of Machalová et al. (2016).

To set the notation, call $\Delta\lambda := \{\lambda_0 = a < \lambda_1 < \ldots < \lambda_g < b = \lambda_{g+1}\}$ a given sequence of knots, and denote by $\mathcal{S}_k^{\Delta\lambda}[a, b]$ the vector space of polynomial splines of degree $k > 0$, defined on I given the knots $\Delta\lambda$. Every spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$ has a unique representation as (see de Boor (1978), Dierckx (1993) for details)

$$s_k(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x). \tag{12}$$

Here, one needs additional knots to build all basis functions of $\mathcal{S}_k^{\Delta\lambda}[a, b]$. Indeed, given the sequence of knots $\Delta\lambda$, one can only build $g-k+1$ linearly independent B-splines of degree k (see, e.g., Dierckx (1993), p.4). To obtain a complete basis for the vector space $\mathcal{S}_k^{\Delta\lambda}[a, b]$, one needs a set of $2k$ linearly independent splines in addition to the latter, B_i^{k+1} for $i = -k, \ldots, -1$ and $i = g-k+1, \ldots, g$, which are precisely generated according to the additional knots. Note that the addition of knots to obtain a complete basis is a common practice, widely documented in

the literature on B-spline bases (e.g., Dierckx (1993), p.10-11 or de Boor (1978), p. 99). Without loss of generality, we here assume that those additional knots are at the boundary, i.e., $\lambda_{-k} = \cdots = \lambda_{-1} = \lambda_0$, $\lambda_{g+1} = \lambda_{g+2} = \cdots = \lambda_{g+k+1}$.

For the purpose of this work, we focus on smoothing splines with zero-integral constraints – i.e., suitable to approximate clr-transformed data. As proved in Machalová et al. (2016), the *optimal* smoothing spline admits a unique representation

$$s_k^i(x) = \sum_{j=-k}^{g} Y_{i,j+k+1} B_j^{k+1}(x), \tag{13}$$

where the vector of B-spline coefficients $\mathbf{Y}_{(i)} = (Y_{i,1}, \ldots, Y_{i,g+k+1})'$ is given by

$$\mathbf{Y}_{(i)} = \mathbf{V}\mathbf{Z}_{(i)}, \quad i = 1, \ldots, n. \tag{14}$$

Here $\mathbf{V}$ is a $(g+k+1) \times D$ matrix which depends only on the position of spline knots and on the possible smoothing parameter (see Appendix C for further details). If the same B-spline basis system is used for all the data, (14) can be expressed in matrix notation as

$$\underline{\mathbf{Y}} \;=\; \underline{\mathbf{Z}}\mathbf{V}', \tag{15}$$

where $\underline{\mathbf{Y}}$, $\underline{\mathbf{Z}}$ are the $N \times (g+k+1)$ matrices

$$\underline{\mathbf{Y}} = \begin{pmatrix} \mathbf{Y}_{(1)} \\ \vdots \\ \mathbf{Y}_{(N)} \end{pmatrix}, \qquad \underline{\mathbf{Z}} = \begin{pmatrix} \mathbf{Z}_{(1)} \\ \vdots \\ \mathbf{Z}_{(N)} \end{pmatrix}.$$

In (Machalová et al., 2016), the explicit expression for the optimal smoothing B-spline is given. As an element of innovation, the following Theorem 5.1 characterizes all the B-splines with zero-integral (not necessarily a smoothing spline), through a necessary and sufficient condition on the vector $\mathbf{b}$ of B-spline coefficients.

Theorem 5.1. *For a spline* $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$, $s_k(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x)$, *the condition* $\int_a^b s_k(x)\,dx = 0$ *is fulfilled if and only if* $\sum_{i=-k}^{g} b_i(\lambda_{i+k+1} - \lambda_i) = 0$.

The proof of Theorem 5.1 is provided in Appendix B. In the light of Theorem 5.1, it is easy to see that vector $\mathbf{b}$ is orthogonal to the vector $\mathbf{n} = (\lambda_1 - \lambda_{-k}, \ldots, \lambda_{g+k+1} - \lambda_g)'$, which only depends on the knots positions. Further, for the vectors $\mathbf{Y}_{(i)}$, $i = 1, ..., n$, of B-spline coefficients, one has the linear constraints

$$\sum_{j=1}^{g+k+1} Y_{ij}(\lambda_j - \lambda_{j-k-1}) = 0. \tag{16}$$

Whenever the same B-spline basis is employed for all the data – as it is usually the case – the linear constraint (16) turns into a model singularity, as we shall show in the next Subsection.

5.2. Regression modeling of B-spline coefficients

By considering the B-spline representations of the clr-transformed response functions $\mathrm{clr}(y_i)(t)$, $i = 1, \ldots, N$, we can express the model (5) in the form of a multivariate regression model. For the purpose of regression modeling, the spline coefficients for the i-th observation $y_i(t)$ are denoted by $\mathbf{Y}_{(i)} = (Y_{i,1}, \ldots, Y_{i,g+k+1})'$, $i = 1, 2, \ldots, N$. Vectors $\mathbf{Y}_{(1)}, ..., \mathbf{Y}_{(N)}$ form the rows of the $N \times g + k + 1$ (random) response matrix $\underline{\mathbf{Y}}$. On this basis, we consider in place of (5) the multivariate linear regression model of the form

$$\underline{\mathbf{Y}}_{(N\times(g+k+1))} = \mathbf{X}_{(N\times p)}\mathbf{B}_{(p\times(g+k+1))} + \underline{\boldsymbol{\varepsilon}}_{(N\times(g+k+1))}, \tag{17}$$

or, equivalently,

$$(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{g+k+1}) = \mathbf{X}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{g+k+1}) + (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_{g+k+1}).$$

Here, the design matrix $\mathbf{X}$ is assumed to be of full column rank, $\boldsymbol{\beta}_j = (\beta_{j0}, \ldots, \beta_{jr})'$, $j = 1, 2, \ldots, g+k+1$, is a vector of unknown regression coefficients and $\underline{\boldsymbol{\varepsilon}}$ is a matrix of random errors. The multivariate responses $\mathbf{Y}_{(i)} = (Y_{1,i}, \ldots, Y_{g+k+1,i})'$, $i = 1, 2, \ldots, N$, are independent with the same unknown variance-covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathrm{cov}(\mathbf{Y}_{(i)}, \mathbf{Y}_{(j)}) = \mathbf{0}_{((g+k+1)\times(g+k+1))}$, $i \neq j$, $\mathrm{var}(\mathbf{Y}_{(i)}) = \boldsymbol{\Sigma}_{((g+k+1)\times(g+k+1))}$, for $i = 1, \ldots N$.

The best linear unbiased estimator (BLUE) of the parameter matrix $\mathbf{B}$ is found as

$$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{g+k+1}), \tag{18}$$

which is invariant to $\boldsymbol{\Sigma}$. Under the assumption that $\underline{\mathbf{Y}}$ is of full column rank, the multivariate model can be simply decomposed into $g+k+1$ univariate multiple regression models that implies an alternative estimation of columns of $\mathbf{B}$ as

$$\widehat{\boldsymbol{\beta}}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_j, \ j = 1, \ldots, g+k+1. \tag{19}$$

The variance-covariance matrix of the vector $\mathrm{vec}(\widehat{\mathbf{B}}) = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2', \ldots, \widehat{\boldsymbol{\beta}}_{g+k+1}')'$ is

$$\mathrm{var}\left[\mathrm{vec}(\widehat{\mathbf{B}})\right] = \boldsymbol{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1},$$

where the symbol $\otimes$ denotes the Kronecker product. The unbiased estimator of $\boldsymbol{\Sigma}$ is $\widehat{\boldsymbol{\Sigma}} = \underline{\mathbf{Y}}'\mathbf{M}_{\mathbf{X}}\underline{\mathbf{Y}}/(n-p)$, where $\mathbf{M}_{\mathbf{X}} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a projector on the orthogonal complement of the vector space $\mathcal{M}(\mathbf{X})$ generated by the columns of the matrix $\mathbf{X}$, i.e., $\mathcal{M}(\mathbf{X}) = \{\mathbf{X}\mathbf{u} : \mathbf{u} \in \mathbb{R}^p\}$.

As the realization of multivariate response $\mathbf{Y}_{(i)}$ is the vector of B-spline coefficients $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ of the clr-transformed data, the variables $Y_{i,1}, \ldots, Y_{i,g+k+1}$ are linearly dependent. Indeed, one has that $\sum_{j=1}^{g+k+1} Y_{ij}(\lambda_j - \lambda_{j-k-1}) = 0$, due to Theorem 5.1. Accordingly, one may expect that a similar constraint applies to the corresponding estimated coefficients, as stated by the following result.

Proposition 5.2. *If* $\sum_{j=1}^{g+k+1} Y_{ij}(\lambda_j - \lambda_{j-k-1}) = 0$ *for all* $i = 1 \ldots, n$, *then* $\sum_{j=1}^{g+k+1} \widehat{\beta}_{sj}(\lambda_j - \lambda_{j-k-1}) = 0$ *for all* $s = 0, \ldots, r$.

The latter constraint introduces a singularity into the regression model (17), which may affect parameter inference. Similarly as in univariate regression (e.g., Fišerová et al., 2007), the model singularity may be an issue when statistical inference is performed based on B-spline coefficients, e.g., when testing for the significance of the coefficient $\boldsymbol{\beta}_j$ through parametric tests based on Fisher's statistics. In these cases, orthonormal representations of the B-spline coefficients may be considered, in the light of the results of Section 5.1. Indeed, vectors $\mathbf{Y}_{(i)}$, $i = 1, ..., N$, form a hyperplane $\mathcal{H}$ of dimension $g + k$, orthogonal to the normal vector $\mathbf{n} = (\lambda_1 - \lambda_{-k}, \ldots, \lambda_{g+k+1} - \lambda_g)'$. One may build an orthonormal basis for $\mathcal{H}$, express $\mathbf{Y}_{(i)}$, $i = 1, ..., N$, through the coordinates of such a basis – removing the singularity due to the linear constraints induced by (9) – and then use the regularized representation for the purpose of further statistical inference. A basis for $\mathcal{H}$ can be easily obtained as the set of the first $g + k$ principal components of the B-spline coefficient vector, which in turn correspond to the Simplicial Functional Principal Components (SFPCs) of the smoothed densities $y_1(t), ..., y_N(t)$ (Hron et al., 2016). However, the BLUE estimation (18) of the regression coefficients is not affected by the singularity constraint in the response, and can be thus computed explicitly, without resorting to the SFPCA or to orthonormalized representations.

It should be also noted that the number of knots for the B-spline basis function cannot be chosen independently of the discretization used to build vectors $\mathbf{W}_i = (W_{i1}, \ldots, W_{iD})'$, $i = 1, ..., N$ (i.e., the discrete compositions which form the raw data). The number of knots and the number of classes on I upon which $\mathbf{W}_i$ are built are indeed related, as the former cannot exceed the latter. The problem of setting the discretization on I and the number of knots is affected from the bias-variance trade-off. Indeed, when building $\mathbf{W}_i$, a fine discretization yields minimum bias in estimating the point value of the target density, but inflates the associated variance, and vice-versa. Similarly in the B-spline representation – where the number of knots is concerned – a high number of knots is associated with low bias and high variance, and vice-versa. Clearly, no optimal choice is known *a priori* to set these parameters, but the 'optimum' is problem dependent. For instance, it depends on the sample size, as well as on the signal-to-noise ratio. Several methods have been developed in the theory of descriptive statistics to set an optimal number of classes when building a histogram. Amongst these, we mention the Sturges' rule (Sturges, 1926), which will be used in the case study of Section 8. Fixed the discretization, the number of knots can be then set as to balance the fitting to the raw data and the smoothness of the estimates, possibly based on a cross-validation analysis as in the case study here presented.

6. Smoothing splines and regression: the relation with the multivariate setting

A natural question which may arise in the proposed context regards the smoothing properties of the regression estimates, and particularly if and how

the data smoothing reflects on the estimates. The key point that we here aim to investigate is whether equivalence results can be stated for the following alternative procedures: (a) the data are smoothed and the Bayes space regression of Section 4 is applied (hereafter named "regression-smoothing"), and (b) a compositional regression (Egozcue et al., 2012) is applied, the model

$$\mathbf{Z}_i \;=\; \boldsymbol{\beta}_0^{(Z)} + \sum_{j=1}^{r} \boldsymbol{\beta}_j^{(Z)} x_{ij} + \boldsymbol{\epsilon}_i, \tag{20}$$

is estimated and the estimates (or predictions) of $\underline{\mathbf{Z}}$ are smoothed afterward (hereafter named "smoothing-regression"). In particular, we here show that, under specific conditions, the following scheme represents the relation between the model here presented and that proposed in (Egozcue et al., 2012)

$$\begin{array}{ccc} \mathbf{Z} & \xrightarrow{\;smoothing\;} & \underline{\mathbf{Y}} \\ {\scriptstyle regression}\downarrow & & \downarrow{\scriptstyle regression} \\ \widehat{\mathbf{Z}} & \xrightarrow[\;smoothing\;]{} & \widehat{\underline{\mathbf{Y}}} \end{array} \tag{21}$$

From (18), the matrix of predicted coefficients $\underline{\mathbf{Y}}$ is obtained as

$$\widehat{\underline{\mathbf{Y}}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Y}}, \tag{22}$$

while for model (20) one has

$$\widehat{\underline{\mathbf{Z}}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Z}}. \tag{23}$$

Plugging-in (15) in (22) we obtain $\widehat{\underline{\mathbf{Y}}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Z}}\mathbf{V}'$. On the other hand, when smoothing splines for $\widehat{\mathbf{Z}}_i$, $i = 1, \ldots, n$, are considered, the matrix of the corresponding B-spline coefficients is obtained as

$$\widehat{\widehat{\underline{\mathbf{Y}}}} \;=\; \widehat{\underline{\mathbf{Z}}}\mathbf{V}_Z'. \tag{24}$$

In order to guarantee that $\mathbf{V}_Z$ coincides with the matrix $\mathbf{V}$ in (15), one needs to build the smoothing spline upon the same sequence of knots, the same degree of spline and the same objective functional (e.g., the same penalization). In this case, and using (23), the matrix $\widehat{\widehat{\underline{\mathbf{Y}}}}$ can be written in the form

$$\widehat{\widehat{\underline{\mathbf{Y}}}} \;=\; \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Z}}\mathbf{V}',$$

that directly implies the target assertion, i.e., $\widehat{\widehat{\underline{\mathbf{Y}}}} \;=\; \widehat{\underline{\mathbf{Y}}}$. As a consequence, when smoothing splines are considered, the smoothness of the observations induces a corresponding degree of smoothness on the estimates, even if this is not explicitly imposed through the use of a PENSSE criterion as that introduced in Section 3.

It should be noted that, although under particular conditions the "smoothing-regression" and "regression-smoothing" approaches are equivalent, the proposed framework provides a more flexible setting to perform the analysis. For instance, to carry out the analysis in the "regression-smoothing" setting, one would need to estimate all the histograms according the same set of classes, which may not be optimal for all of them. In the "smoothing-regression" setting, one can freely estimate the histograms with their own optimal classes and then fit the basis expansion to each of those. In other cases, one may be already provided with densities defined over a fine grid (e.g., with particle-size data, Menafoglio et al. (2014, 2016a,b)). Dealing with high-dimensional (compositional) data from a discrete viewpoint may yield issues related with the curse of dimensionality, which are completely overcome with a functional viewpoint.

7. A Simulation study

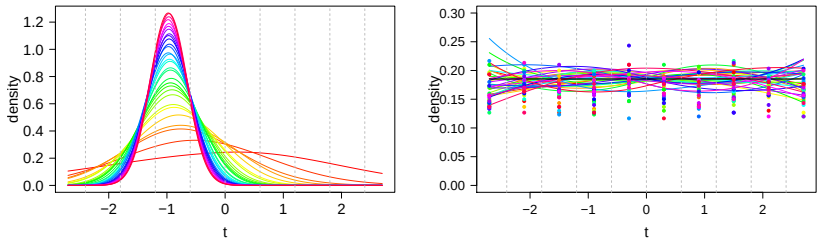
7.1. Assessing the effects of smoothing on regression

A simulation study is here presented to test the performances of the proposed methodology. Attention will be paid to the sensitivity of the results to the main parameters – number of classes, of knots and of starting data. To generate the functional dataset, $y_i \in \mathcal{B}^2, i = 1, ..., N = 30$, the following reference model was considered,

$$y_i(t) = \beta_0(t) \oplus [x_i \odot \beta_1](t) \oplus \varepsilon_i(t), \quad t \in I = [-3, 3], \quad i = 1, \ldots, 30, \tag{25}$$

where ε is the random error, whose mean is the neutral element of perturbation in $\mathcal{B}^2$, i.e., the uniform distribution. Specifically, for each observation i, 500 realizations were generated from a uniform distribution on $I = [-3, 3]$ and then smoothed to represent the errors ε_i through a B-spline basis. The smoothing procedure was designed to reproduce the estimation strategy which will be applied in the case study of Section 8, and was based on quadratic splines ($k = 2, l = 1$), with smoothing parameter $\alpha = 0.99$ and five equally spaced knots. The regression parameters were set to truncated Gaussian densities $N(\mu_i, \sigma_i^2)$, $i = 0, 1$, with support on I. For the intercept β_0, the parameters were set to $\mu_0 = 0, \sigma_0 = 2$, and, for the slope parameter β_1, to $\mu_1 = -1, \sigma_1 = 1$. For brevity, the latter model is hereafter named *Model 500*, 500 indicating the number of sampled data.

To test the robustness of the method to the number D of classes upon which histogram data are built, the model was estimated for three different parameter settings, one of which determined by using the Sturges' rule, the others by a higher/lower number of classes. Specifically, with the previous model settings, the Sturges' rule suggested an optimal value of $D = 10$ classes. The two additional values considered were $D = 7$ and $D = 14$. For each set of parameters, the functional dataset was generated from the model (25); the regressors x_i were set to 30 equally spaced values in the interval $[0.01, 10]$. To assess the performances of the method, the simulation was replicated $K = 30$ times. Figure 1 represents the observed response for the first sample, together



(a) Reference response. (b) Smoothed errors via B-splines.

Figure 1: Dataset used for the simulation study: simulated (true) density response and model errors. Vertical dashed grey lines indicate the partition of the interval I in 10 classes.

with raw and smoothed error model data, with $D = 10$ classes and 3 knots. To compare the quality of the obtained estimates, the integrated square error (ISE) between the true and estimated density parameter functions, $ISE = \|\beta_{l_k} \ominus \hat{\beta}_{l_k}\|^2_{\mathcal{B}}$, $l = 0, 1$, $k = 1, ..., 30$, was considered. The top panels of Figure 2 display the boxplots of the integrated square errors for $D = 7, 10, 14$. Simulations show that the Sturges' rule can be considered as a reasonable choice, since the estimates of both coefficients do not appear to be sensitive to that parameter setting.

Having fixed the number of classes according to Sturges' rule, sensitivity of the result to the number of knots was assessed in the same simulation setting, only varying the number of equally spaced knots in $\{3, 5, 7\}$. The bottom panels of Figure 2 seem to suggest the use of a moderate number of knots, as higher number of knots may lead to overfitting the data. Note that the parameters β_0, β_1 are two-dimensional in the Bayes space (they belong to an affine space of dimension 2, Hron et al. (2016)), compatible with the low number of knots suggested by simulation results.

Finally the experiment was repeated with different numerosity of the initial sampled error data, taking 300 (*Model 300*) and 700 (*Model 700*) of them. Results are consistent with the previous ones, hence omitted. They confirm the overall good performances of Sturges' rule, and suggest a moderate number of classes in all the cases. In particular, they suggest that the number of sampled data does not have a strong influence on the results (see Figure 3). Note that all the simulation settings here considered are based upon a relatively high number of data, as our proposal is deemed to address this situation rather than the case of small datasets (see also the discussion in Section 9).

7.2. Comparison of the Bayes approach with competitors in L^2

In this subsection, the proposed approach is compared with two alternative methods to fit a linear model based on the same distributional responses $\{y_i, i = 1, ..., 30\}$ (Figure 1) and the same scalar regressors $\{x_i, i = 1, ..., 30\}$

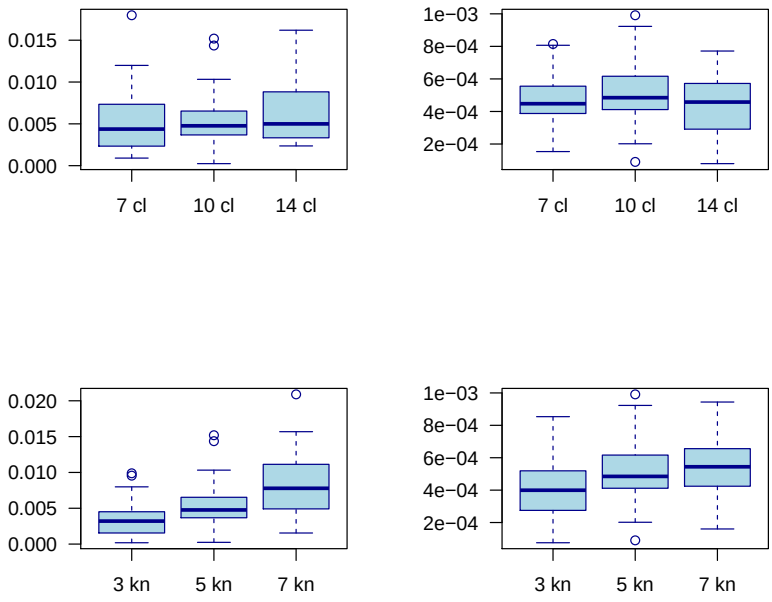


Figure 2: Boxplots of ISE between the true density parameter functions β_0 (left) and β_1 (right) and their estimates. The boxplots are built upon $K = 30$ repetition of the estimation of Model 500. Top panels: sensitivity to the number of classes (cl.): 7, 10 (Sturges' rule) and 14. Bottom panels: sensitivity to the number of knots in $\{3, 5, 7\}$.

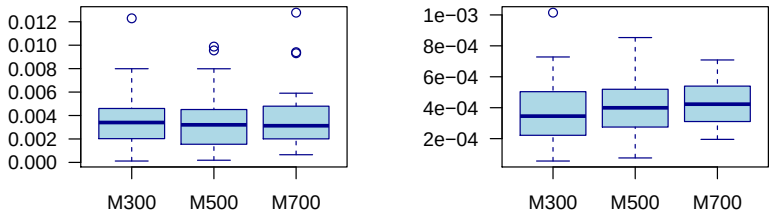


Figure 3: Comparison of the boxplots of ISE for models with different number of starting data with optimal parameter setting – Model 300 (9 classes, 5 knots), Model 500 (10 classes, 3 knots), Model 700 (10 classes, 5 knots) – for β_0 (left) and β_1 (right).

as in Subsection 7.1. To this end, the following models are considered: (a) a function-on-scalar model in L^2

$$y_i(t) = \alpha_0(t) + \alpha_1(t)x_i + \zeta_i(t), \tag{26}$$

with ζ_i random errors in L^2, with zero mean; and (b) a function-on-scalar model in L^2, but for the log-transformation of the response

$$\log(y_i(t)) = \gamma_0(t) + \gamma_1(t)x_i + \eta_i(t), \tag{27}$$

with η_i random errors in L^2, with zero mean. Note that using a logarithmic transformation is very common for data on a relative scale, and preserves positiveness, but does not guarantee that the resulting estimates integrate to unity. Estimation of the models (26) and (27) was obtained by ordinary least squares, and computed numerically on a fine discretization of the data.

Note that the regression coefficients of the proposed model and of the alternative ones cannot be directly compared, and so their estimates. Thus, the results of the three methods were compared in terms of (i) goodness of fit on the (simulated) response in Figure 4a and (ii) quality of predictions in correspondence of 20 equally spaced new values of the regressors x in $[0, 30]$ (Figure 4b).

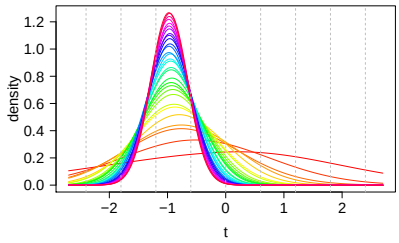
Figure 4 shows the results obtained when using the Bayes space methodology. In particular, it shows that the proposed method reproduces precisely the parameters $\boldsymbol{\beta}$ generating the model (Figure 4e).

In fact, very different results are obtained when using the geometry of L^2 (Figure 5). The model in L^2 clearly provides poorer estimations since fitted (Figure 5a) as well as predicted responses (Figure 5b) do not follow the integral constraint and exhibit negative values. Moreover, the difference between predicted curves in L^2 and in $\mathcal{B}^2$ is evident, the latter being much more precise in representing the reference realizations. Here, predictions have greater variance around the mean for increasing values of the regressor x and they are more elongated in their amplitude which is also a consequence of analysing densities on an absolute scale.

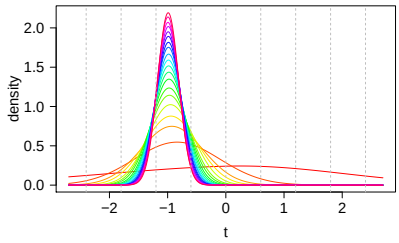
Using a log-transformation allows to improve the results both in terms of fitting and of prediction with respect to an L^2 approach (see Figure 6), as the resulting densities are guaranteed to be positive. However, fitted and predicted responses do not honor the unit integral constraint, thus providing unsatisfactory results.

8. Modeling metabolite distributions in newborns

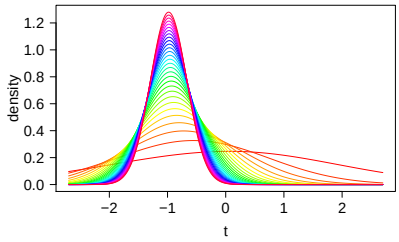
The data used in this example are part of a standard newborn screening done in 2013 in the Laboratory of Inherited Metabolic Disorders, in the Department of Clinical Biochemistry of the Faculty Hospital in Olomouc. Here, the weight and gender of every newborn are observed, together with 48 metabolic parameters (so called metabolites) measured from dried blood spots of each newborn.



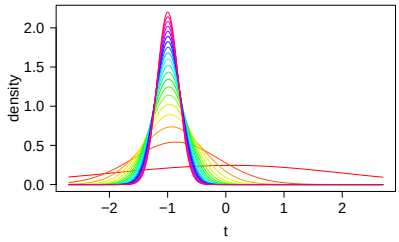
(a) Simulated response.



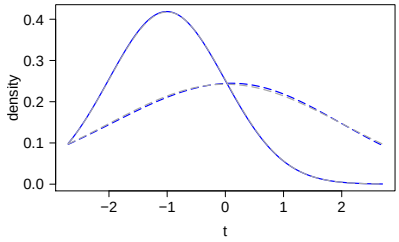
(b) Target realization to be predicted.



(c) Fitted response y, in $\mathcal{B}^2$.

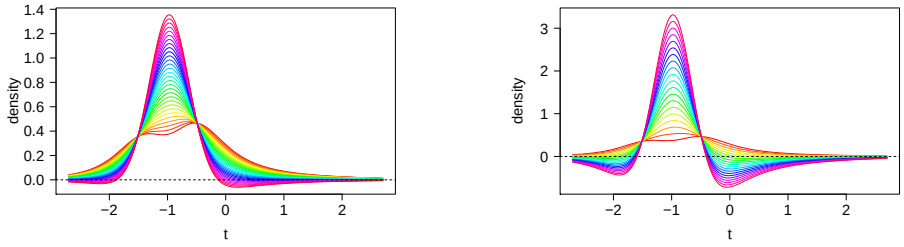


(d) Predicted response y, in $\mathcal{B}^2$.



(e) True (grey line) and estimated (blue line) parameters.

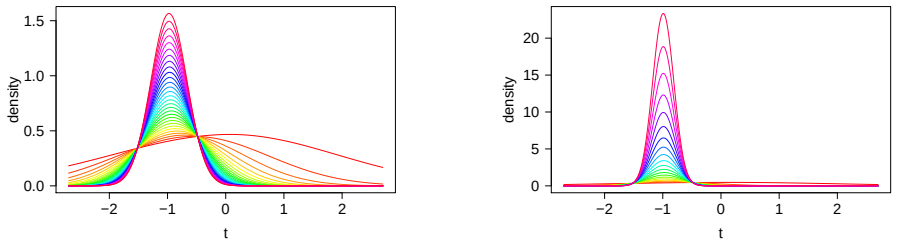
Figure 4: Fitted response and estimated parameters in $\mathcal{B}^2$ space.



(a) Fitted response y, in L^2.

(b) Predicted response y, in L^2.

Figure 5: Fitted and predicted model in L^2 space.



(a) Fitted response y, using log-transformation.

(b) Predicted response y, using log-transformation.

Figure 6: Fitted and predicted model in L^2 space using log-transformation.

The dataset we consider collects the data about 10000 newborns with standard weights. All the data were anonymised prior to analysis, and were not yet used elsewhere. Although they are not publicly available, their aggregations are given in Appendix D. In particular, for the purpose of this example, we focus on the metabolite C18, which is presumed to be closely connected with the weight of newborns. More in general, newborn screening is a nationwide active search of diseases in their early, preclinical stage, so that these diseases are diagnosed and treated before they may impact a child and cause irreversible health damage. The screening is based on the analysis of dried blood spots on filter paper; blood is taken under defined conditions for all newborns born in the Czech Republic and 18 diseases are investigated.

For the purpose of modeling the dependence of C18 distribution on weight through functional regression models, the C18 distribution was assessed from sampled data as follows. The values of the logarithm of C18 were divided into 10 groups of equal size according to the logarithm of weight, and represented by the midpoint of the corresponding interval of weights, separately for girls (g) and boys (b). In order to exclude extreme values of concentration of the metabolite, the measurements under the bottom 0.5%-quantile and above the upper 99.5%-quantile were omitted. In each of the 10 groups, the distribution of log(C18) was estimated empirically, by dividing in equally-spaced classes and computing the frequency within each class. Here, the number of optimal classes was computed by using Sturges' rule, resulting in 9.93 for girls and 9.94 for boys. Hence, for both girls and boys we built $D = 10$ equally-spaced classes on the ranges $I_g = [-2.936, -0.939]$ and $I_b = [-2.813, -0.763]$. Tables 2 and 3 in Appendix D list the vectors of proportions $\mathbf{W}_i = (W_{i1}, \ldots, W_{i10})', i = 1, \ldots, 10$, of log(C18) within each group of weights, together with the midpoints of the classes $t_j, j = 1, \ldots, 10$. On these bases, the vectors of proportions were transformed by using the discrete version of the clr transformation, resulting in the vectors $\mathbf{Z}_i$ with element values $Z_{i1}, \ldots, Z_{i10}, i = 1, \ldots, 10$, reported in Tables 4 and 5 of Appendix D. Note in particular the condition $\sum_{j=1}^{10} Z_{ij} = 0$ for each $i = 1, \ldots, 10$.

As a second step of the analysis, the clr-transformed proportions were smoothed by using a system of smoothing splines with support I_g and I_b, for girls and boys, respectively, fulfilling the zero-integral constraint, as described in Section 5.1. In both cases (i.e., for girls and boys) the same strategy was followed to set the values of the parameters. We considered quadratic splines (i.e., $k = 2$, $l = 1$) and set the number of knots by performing leave-one-out cross-validation. The latter showed that the results are robust to the number of knots in the set $\{3, 5, 7\}$. The optimal smoothing spline $s_k(t)$ on I was found as to minimize the penalized functional

$$J_l(s_k) = (1 - \alpha) \int_I \left[s_k^{(l)}(s) \right]^2 ds + \alpha \sum_{j=1}^{D} w_j^s \left[Z_{ij} - s_k(t_j) \right]^2,$$

where the parameter α was set to $\alpha = 0.99$ in order to be as close as possible to data (t_j, Z_{ij}), and the weights were set to $w_j^s = 1$, for $j = 1, \ldots, 10$. The

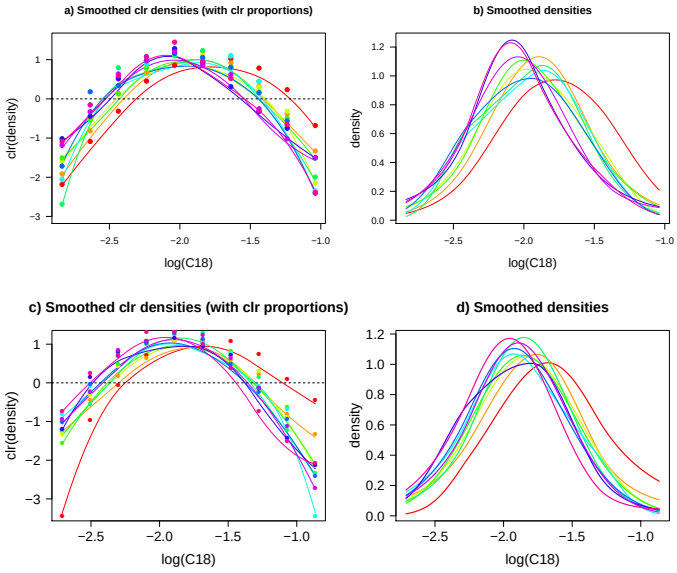


Figure 7: Clr densities and their inverse (i.e., the densities) of log(C18). Girls a)-b), boys c)-d).

resulting smoothed clr-densities $y_{c_i}(t) \in L^2(I)$, $i = 1, ..., 10$, are displayed in Figure 7 together with the corresponding densities $y_i(t) \in \mathcal{B}^2(I)$, $i = 1, ..., 10$, obtained by applying the inverse clr transformation to the smoothed data, i.e., $y_i(t) = \text{clr}^{-1}[y_{c_i}](t) = \mathcal{C}[\exp(y_{c_i})](t), i = 1, \ldots, 10, t \in I$.

Given that the supports of the log(C18) distribution differ between girls and boys populations, for each of the two groups, we separately modeled the dependence of the log(C18) distributions on log(weight) through following linear model in $\mathcal{B}^2(I)$,

$$y_i(t) = \beta_0(t) \oplus [\log(w_i) \odot \beta_1](t) \oplus \varepsilon_i(t), \quad i = 1, \ldots, 10. \tag{28}$$

By considering the same B-spline basis functions $B^3_{-2}(t), \ldots, B^3_3(t)$ for the response $\text{clr}(y(t))$, the regression parameters $\text{clr}[\beta_0(t)]$, $\text{clr}[\beta_1(t)]$ and the error $\text{clr}[\varepsilon(t)]$, the model (28) can be written as a multivariate model for the B-spline coefficients $Y_{i1}, \ldots, Y_{i6}$, in matrix form as $\underline{\mathbf{Y}} = \mathbf{XB} + \underline{\boldsymbol{\varepsilon}}$. The resulting estimates $\widehat{\boldsymbol{\beta}}_0 = (\widehat{\beta}_{01}, \widehat{\beta}_{02}, \cdots, \widehat{\beta}_{06})'$ and $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\beta}_{11}, \widehat{\beta}_{16}, \cdots, \widehat{\beta}_{16})'$ for girls and boys are listed in Table 1, together with the estimates of their standard deviations. The corresponding estimates of the regression functions $\text{clr}[\beta_0(t)]$ and $\text{clr}[\beta_1(t)]$ are displayed in Figure 8, together with their counterparts in $\mathcal{B}^2(I)$. Here, the colors distinguish the gender – red for girls and blue for boys.

We first focus on the interpretation of the estimated regression parameters in the female group, by visual inspection of Figure 8 (top panels). The intercept $\beta_0(t)$ is hardly interpretable, as it estimates the expected value of the density of

	Estimates of regression parameters $\beta_{\cdot 1}, \ldots, \beta_{\cdot 6}$					
$\widehat{\boldsymbol{\beta}}_0^g$	-17.693	-14.437	-9.227	7.573	17.487	16.553
$\widehat{\sigma}$	7.491	5.995	3.235	3.436	3.998	7.536
$\widehat{\boldsymbol{\beta}}_1^g$	1.978	1.738	1.265	-0.835	-2.235	-2.274
$\widehat{\sigma}$	0.928	0.742	0.403	0.425	0.495	0.933
$\widehat{\boldsymbol{\beta}}_0^b$	-33.132	-13.687	-7.866	5.601	21.190	24.920
$\widehat{\sigma}$	6.828	3.054	2.028	1.984	4.572	9.292
$\widehat{\boldsymbol{\beta}}_1^b$	3.912	1.660	1.105	-0.585	-2.727	-3.337
$\widehat{\sigma}$	0.841	0.377	0.245	0.249	0.563	1.145

Table 1: Estimates of regression parameter vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ with marking – g for girls, b for boys (colourless rows), together with the corresponding estimates of the standard deviations $\widehat{\sigma} = \left\{\widehat{\text{var}}(\text{vec}(\widehat{\mathbf{B}}))\right\}_{k,k}$ (grey rows).

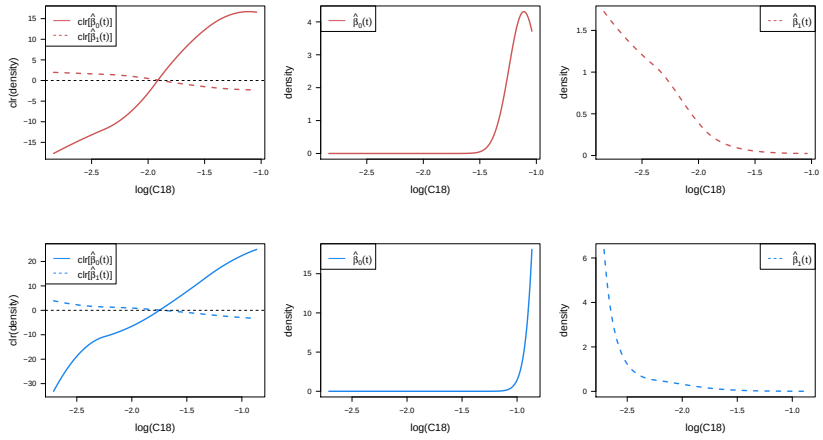


Figure 8: Estimates of regression coefficients. Upper plots represent the results for the girls' group, lower plots those for the boys' group. In both lines, from left to right: estimates in L^2 (clr-transformed), estimate of β_0 in $\mathcal{B}^2$, estimate of β_1 in $\mathcal{B}^2$.

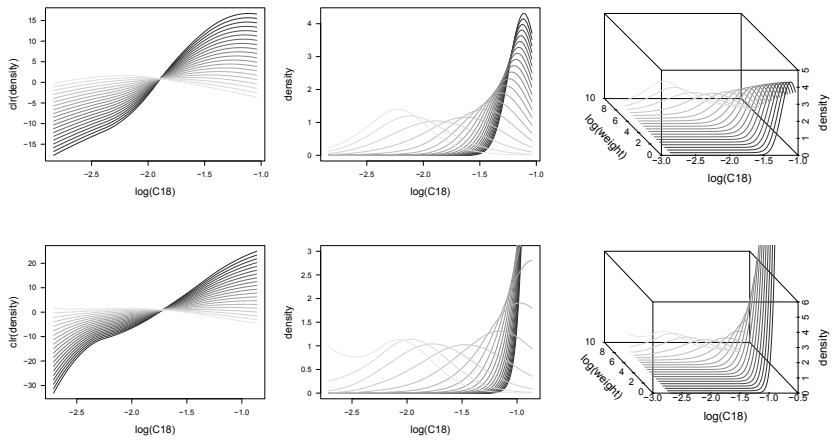


Figure 9: 2D and 3D graphs of predicted distributions for increasing sequence of 20 values of log weights.

log(C18) when the weight of a newborn is 1 gram. Nevertheless, the coefficient $\beta_0(t)$ acts as a shift in the model – in sense of geometry of $\mathcal{B}^2$ – towards a density highly concentrated in the right tail of the domain I_g. Instead, by graphical inspection of the same figure, one can better interpret the effects of the slope coefficient $\beta_1(t)$ on the response. Indeed, if the weight of newborns increases, the predicted average distribution of log(C18) tends to be more concentrated in the left part of domain I_g, and viceversa. This can be better appreciated from Figure 9, where the response $y(t)$ is predicted for a sequence of increasing values of the log-weights in the interval $[\log(w_1), \log(w_{20})] = [\log(1), \log(7000)]$. Note that, as the value of the regressor increases, the predicted expected values of the log(C18) decreases while its predicted variance increases. It can be concluded that relative proportion of newborns with higher concentrations of metabolite C18 decreases when weight increases, while the relative proportion of newborns with middle and lower concentrations of C18 increases. In general, newborns with lower weight exhibit higher concentrations of metabolite C18 whereas those with higher weight show middle and lower concentrations of C18. Very similar conclusions can be drawn for the males' group. However, here the impact of lower weight to metabolite distribution seems to be even more dramatic. This indicates that the impact of the underweight to the predisposition to a metabolic disease is even more serious for boys than for girls.

The fitted curves corresponding to the $N = 10$ observed distributions are displayed in Figure 10 with the same gender color scheme. To assess the goodness-of-fit of the model on the observed density curves, a pointwise version of coefficient of determination $R^2(t)$, $t \in I$, was computed based on the pointwise comparison between the predicted clr-transformed density and the actual data. Additionally, a global coefficient of determination, denoted by R^2_{glob}, was com-

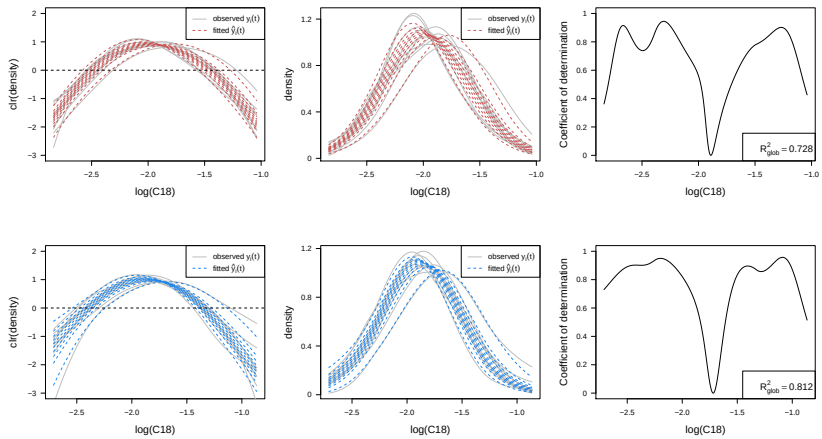


Figure 10: Comparison of observed y (grey) and fitted $\hat{y}$ (colored curves) distributions in L^2 and $\mathcal{B}^2$ (left and central panels). Pointwise coefficient of determination (right panels, upper for girls and lower for boys).

puted as

$$R^2_{glob} = \frac{\sum_{i=1}^{N} \|\mathrm{clr}(\hat{y}_i) - \mathrm{clr}(\bar{y})\|_2^2}{\sum_{i=1}^{N} \|\mathrm{clr}(y_i) - \mathrm{clr}(\bar{y})\|_2^2}.$$

The latter measures the amount of the total sample variance of the $y_i(t)$ explained by the model, in a global sense. The pointwise and the global coefficients of determination are displayed in Figure 10. Although the graphs of the pointwise R^2 indicate some lack of fit in the central part of the domain, the coefficient R^2_{glob} reaches high values in both cases, being about 72.8% and 81.2%, thus indicating a very good (global) fit of the model.

In order to support the interpretation of the parameters of the regression models, it is desirable to incorporate uncertainty in estimation of regression parameters. To this end, we employed a resampling method (bootstrap), to avoid introducing strong distributional assumptions, such as Gaussianity. In particular, we considered a bootstrap scheme based on re-sampling of the model-residuals. More precisely, having estimated the model, we computed the estimated residuals as $\mathrm{clr}(\hat{\varepsilon}_i) = \mathrm{clr}(y_i) - \mathrm{clr}(\hat{y}_i)$. For each bootstrap repetition, we generated the bootstrap sample $\mathrm{clr}(\varepsilon_1^{boot}), \ldots, \mathrm{clr}(\varepsilon_N^{boot})$ by sampling with repetition from $\{\mathrm{clr}(\hat{\varepsilon}_1), \ldots, \mathrm{clr}(\hat{\varepsilon}_N)\}$. We defined the corresponding bootstrap response variables

$$\mathrm{clr}(y_i^{boot})(t) = \mathrm{clr}(\beta_0)(t) + \log(w_i^{boot}) \cdot \mathrm{clr}(\beta_1)(t) + \mathrm{clr}(\varepsilon_i^{boot})(t), \quad i = 1, \ldots, N,$$

and collect the bootstrap sample

$$S = \left[(\log(w_1^{boot}), \mathrm{clr}(y_1^{boot})), \ldots, (\log(w_N^{boot}), \mathrm{clr}(y_N^{boot})) \right].$$

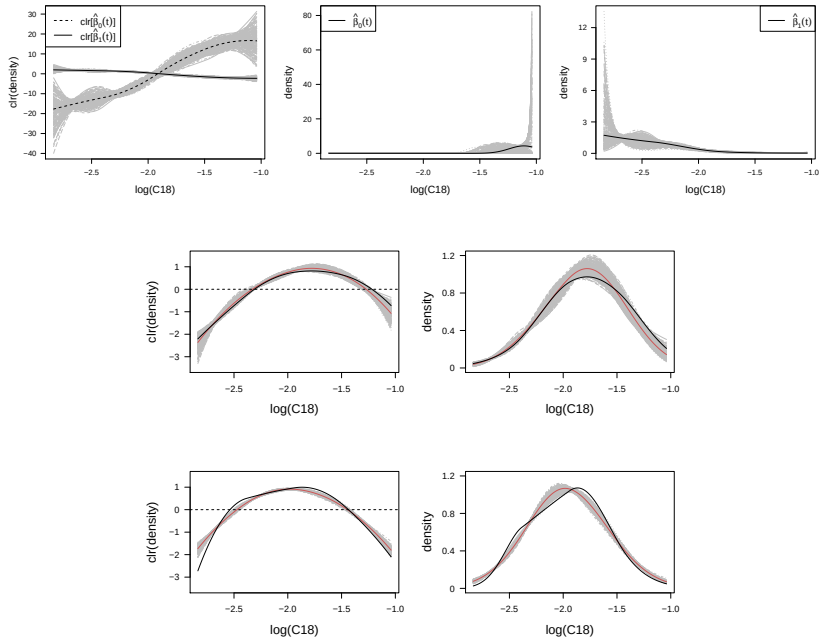


Figure 11: Bootstrap results for the girls' group. Upper three panels: the black curves indicate the estimates of the regression parameters, the grey curves indicate their $R = 200$ bootstrap estimates. Bottom four panels: the black curves indicate observed distributions for w_1 (upper panels) and w_5 (bottom panels), the red curves indicate the fitted distribution for w_1 and w_5 by model (28), the grey lines indicate the corresponding fitted distributions obtained by the bootstrap procedure.

We considered $R = 200$ bootstrap repetitions, which seemed sufficient for the purpose of uncertainty assessment. For each bootstrap sample, we fitted the model and obtained the corresponding estimates of the parameters, denoted by $(\widehat{\beta}_{0r}^{boot}, \widehat{\beta}_{1r}^{boot})$, for $r = 1, ..., R$. The estimated β's and the bootstrap repetitions are displayed in Figure 11.

We then used these bootstrap outputs $(\widehat{\beta}_{0r}^{boot}, \widehat{\beta}_{1r}^{boot})_{r=1,...,R}$ to quantify the uncertainty in the fitted model for fixed values of $\log(w)$. Here, two values of weights were chosen to compute 200 fitted curves by using the estimates obtained by the bootstrap procedure. The results are displayed in Figure 11 for the female group. Similar results are obtained for the male group (not shown for the sake of brevity). Interesting patterns appear in the figure, as most of the uncertainty in β_0 is shown in the right part of the domain, whereas for β_1 it is mostly present in the left part of the domain. The bottom two panels of Figure 11 indicate a poor fitting for the observed distribution corresponding to $\log(w_5)$, which can be also read from the pointwise coefficient of determination (see Figure 10). This can indicate that the response might depend on other

regressors, not available in this study.

Finally, a leave-one-out cross-validation analysis was conducted to assess the goodness of the model in terms of prediction performances. The latter were assessed by computing the mean squared error

$$MSE_{CV} = \frac{\frac{1}{N}\sum_{i=1}^{N} \|y_i - \widehat{y}_i^{(-i)}\|_{\mathcal{B}}^2}{\frac{1}{N}\sum_{i=1}^{N} \|y_i\|_{\mathcal{B}}^2},$$

where $\widehat{y}_i^{(-i)}$ indicates the prediction of the i-th density, by using all the data but the i-th. Results showed that the predictions of the model are satisfactory, with a mean squared error of 4.51% for females and 5.02% for males. Results with a different number of knots for the B-spline basis (namely 3,5,7) were not significantly different (females: 4.16%, 5.17%; males: 5.94%, 4.94%; for 3,7 knots respectively), confirming the robustness of the method to the choice of these parameters.

9. Conclusion and discussion

In this work, a novel approach was presented to perform functional regression when the response is a density function. The theory of Bayes Hilbert spaces was employed to extend the well-known results of FDA to functional compositional data. Using the Bayes space approach allows accounting for the relative nature of PDFs and the related properties (e.g., scale invariance and relative scale), which may be captured only when Bayes Hilbert spaces are considered.

For the actual estimation of the regression coefficients, an approach based on a B-spline expansion was proposed, properly adapted to deal with density data. Here, a key result was proved to characterize the B-spline expansion of clr-transformed data, which provides a representation of the data constraints in terms of a linear constraint on the B-spline coefficients. The singularity problem induced by the latter constraint motivates further research in the direction of building orthonormal bases in the Bayes space, which would allow expressing its elements through a set of unconstrained coefficients, to be further used for the purpose of, e.g., inference on the coefficients based on functional F-tests.

Alternative approaches could be considered, as those arising from the transformations proposed by Ramsay and Silverman (2002, 2005). For instance, Ramsay and Silverman (2005, Sec. 6.6) propose to represent a PDF f over a domain $I = [t_0, t_1]$ as $f(t) = g(W(t)) = \exp W(t)/[\int_I \exp W(\tau)d\tau]$, t in I, where W is an unconstrained L^2 function, expressed through an expansion over a functional basis (e.g., B-splines). This representation is used by Ramsay and Silverman (2005) to perform data smoothing through penalized maximum likelihood. To develop (penalized) least squares methods for regression (and smoothing), one should consider the inverse transformation g^{-1} expressing W as a function of f. In general, g^{-1} involves both point-wise and differential information of the log-density $\log(f)$, and does not coincide either with the log-transform or with the centred log-ratio (clr) transformation, which have been

explored in Section 7. When the function W is represented via a zero-integral function, g^{-1} simplifies, and it precisely coincides with the clr transformation. This latter framework thus coincides with the one we proposed, the transformation being also associated with the meaningful geometric structure of the Bayes space, which provides a proper mathematical setting for least-square regression. In the general case, the estimate of a regression model based on the transformation of (Ramsay and Silverman, 2005) could be possibly based on penalized maximum likelihood methods. This would require a substantial further research along a different yet potentially interesting line with respect to the scope of the present work. Note however that working with the derivative of the log-density (e.g., Ramsay and Silverman (2002), Sec. 5.4) would require regularity assumptions which may not be met in practice, e.g., in the presence of Laplace distributions.

It should be noted that the proposed methodology is rather devoted to large sample sizes than to small ones. Indeed, the main issue with the sample size is related with the need of estimating the entire response distribution from the data – if it is not given, as in our case study – instead of its first moments. Although demanding, this offers the clear advantage of working with the entire information content that the distribution offers. In the cases in which the sample size is an issue instead, one may resort to multivariate approaches (Egozcue et al., 2012). As an alternative, one may elaborate more refined ways to estimate the response distribution within the classes, possibly based on a moving window for the values of the regressors (or a kernel-based method), rather than dividing the regressors in classes and estimating the density within each class. The latter strategy would entail the use of regression models with correlated residuals, which is indeed possible, but requires an extension of the model which will be the scope of future work.

More in general, the possibility of obtaining estimates of an entire distribution has a great potential from the application viewpoint, as our approach enables one to model not only the relation of the mean and the variance of the response on the regressors, but of all the moments jointly. Nevertheless, still critical appears accounting for the uncertainty associated with the various steps of the analysis, particularly the data pre-processing. Ways to account for the latter uncertainty in the estimation procedure are currently under investigation.

As a way to assess the uncertainty of the estimates, a bootstrap resampling method was considered. On this basis, one could also develop confidence bands for the regressor coefficients, e.g., based on depth measures. On the other hand, the bootstrap resampling procedure together with the measures of goodness-of-fit here proposed may support the model selection, or suggest the introduction of further regressors, as shown in Section 8. Although the proposed theory is still limited to the case of scalar regressors, the approach is entirely general and thus could provide the basis to include more complex regressors (e.g., functional and distributional) into the model. This would be of great relevance from the application view-point and will be the scope of future research.

Acknowledgments

The authors gratefully acknowledge both the support by Czech Science Foundation GA15-06991S, the grant IGA PrF IGA_PrF_2017_019 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc, and the grant COST Action CRoNoS IC1408. The authors are also grateful to the anonymous Referees whose valuable comments contributed to improve this work.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Supplementary material.

The codes that implement the proposed regression methods are available as supplementary material of the present article.

Appendix A: computational aspects of function-on-scalar regression in L^2

Suppose that $y_i(t)$ and $\beta_j(t)$ admit the representations

$$y_i(t) = \sum_{k=1}^{K_y} c_{ik}\varphi_k(t), \quad \beta_j(t) = \sum_{k=1}^{K_\beta} b_{jk}\psi_k(t), \tag{29}$$

in terms of known basis systems $\{\varphi_1, \ldots, \varphi_{K_y}\}$ and $\{\psi_1, \ldots, \psi_{K_\beta}\}$ (e.g., B-spline basis), with coefficients $\{c_{ik}\}$ and $\{b_{jk}\}$. Equivalently, we may express (29) in matrix notation as $\mathbf{y}(t) = \mathbf{C}\boldsymbol{\varphi}(t)$ and $\boldsymbol{\beta}(t) = \mathbf{B}\boldsymbol{\psi}(t)$, where $\mathbf{C}$ and $\mathbf{B}$ are matrices of bases coefficients with dimensions $N \times K_y$ and $p \times K_\beta$, respectively, and $\boldsymbol{\varphi}$, $\boldsymbol{\psi}$ are vectors of basis functions.

If in (29) the same basis system is used for both the y's and the β's (i.e., $K \equiv K_y = K_\beta$, $\varphi_k = \psi_k$, $k = 1, ..., K$), the estimation of functions β_j reduces to find the matrix of coefficients $\mathbf{B}$ by minimizing

$$\begin{aligned}\text{PENSSE}(\boldsymbol{\beta}) = &\int_I \left[\mathbf{C}\boldsymbol{\varphi}(t) - \mathbf{X}\mathbf{B}\boldsymbol{\varphi}(t)\right]' \left[\mathbf{C}\boldsymbol{\varphi}(t) - \mathbf{X}\mathbf{B}\boldsymbol{\varphi}(t)\right] dt \\ &+\lambda \int_I \left[L\mathbf{B}\boldsymbol{\varphi}(s)\right]' \left[L\mathbf{B}\boldsymbol{\varphi}(s)\right] ds.\end{aligned} \tag{30}$$

Note that setting $\lambda = 0$ yields the reformulation of (3) in terms of basis expansion.

Further, denote by $\mathbf{P}$, $\mathbf{Q}$ the symmetric constant matrices of order K, $\mathbf{P} = \int_I \left[L\boldsymbol{\varphi}(s)\right] \left[L\boldsymbol{\varphi}(s)\right]' ds$ and $\mathbf{Q} = \int_I \boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)' dt$. By differentiating (30) with respect to $\mathbf{B}$ it can be shown that the estimation of $\mathbf{B}$ is found as solution of the linear system

$$\left(\mathbf{X}'\mathbf{X}\mathbf{B}\mathbf{Q} + \lambda\mathbf{B}\mathbf{P}\right) = \mathbf{X}'\mathbf{C}\mathbf{Q}. \tag{31}$$

System (31) can be equivalently reformulated using the Kronecker product $\otimes$ as

$$\left[\mathbf{Q} \otimes \left(\mathbf{X}'\mathbf{X}\right) + \mathbf{P} \otimes \lambda \mathbf{I}\right] \mathrm{vec}(\mathbf{B}) = \mathrm{vec}\left(\mathbf{X}'\mathbf{C}\mathbf{Q}\right). \tag{32}$$

The matrix $\mathbf{B}$ is thus obtained as the solution of a system of linear equations of dimension $p \times K$.

Appendix B: proofs of theorems

Proof of Theorem 5.1

In the following the notation $s_k^{\mathbf{b}}(x)$ is used to emphasize the dependency on the vector $\mathbf{b} = (b_{-k}, \ldots, b_g)'$. It is known that

$$\int_a^b s_k^{\mathbf{b}}(x)\,\mathrm{d}x \;=\; \left[s_{k+1}^{\mathbf{c}}(x)\right]_a^b,$$

for a vector $\mathbf{c}$, which is given by

$$s_k^{\mathbf{b}}(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x) \;=\; \frac{\mathrm{d}}{\mathrm{d}x} \sum_{i=-k-1}^{g} c_i B_i^{k+2}(x) \;=\; \frac{\mathrm{d}}{\mathrm{d}x} s_{k+1}^{\mathbf{c}}(x). \tag{33}$$

The components of vectors $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ and $\mathbf{c} = (c_{-k-1}, \ldots, c_g)'$ satisfy

$$b_i \;=\; (k+1)\frac{c_i - c_{i-1}}{\lambda_{i+k+1} - \lambda_i}, \quad i = -k, \ldots, g,$$

so that

$$c_i \;=\; c_{i-1} + \frac{b_i\left(\lambda_{i+k+1} - \lambda_i\right)}{k+1}, \quad i = -k, \ldots, g.$$

To simplify the notation we set

$$d_i = \frac{k+1}{\lambda_{i+k+1} - \lambda_i}, \quad i = -k, \ldots, g; \tag{34}$$

then

$$c_i \;=\; c_{i-1} + \frac{b_i}{d_i}, \quad i = -k, \ldots, g.$$

From these $g + k + 1$ equations it is easy to see that

$$c_g \;=\; \frac{b_g}{d_g} + \cdots + \frac{b_{-k}}{d_{-k}} + c_{-k-1}. \tag{35}$$

With respect to (33) it is evident that

$$\int_a^b s_k^{\mathbf{b}}(x)\,\mathrm{d}x \;=\; \left[s_{k+1}^{\mathbf{c}}(x)\right]_a^b \;=\; s_{k+1}^{\mathbf{c}}(\lambda_{g+1}) - s_{k+1}^{\mathbf{c}}(\lambda_0), \tag{36}$$

because $a = \lambda_0$, $b = \lambda_{g+1}$. Considering the definition, the properties of B-splines and the above mentioned additional knots, it follows that

$$s_{k+1}^{\mathbf{c}}(\lambda_{g+1}) - s_{k+1}^{\mathbf{c}}(\lambda_0) = c_g - c_{-k-1}. \tag{37}$$

Thus

$$\int_a^b s_k^{\mathbf{b}}(x)\,\mathrm{d}x \;=\; c_g - c_{-k-1}. \tag{38}$$

Now it is clear that for a spline $s_k^{\mathbf{b}}(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$, $s_k^{\mathbf{b}}(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x)$, the condition

$$\int_a^b s_k^{\mathbf{b}}(x)\,\mathrm{d}x \;=\; 0$$

is fulfilled if and only if

$$c_g \;=\; c_{-k-1}.$$

From (35) it follows that

$$c_g \;=\; c_{-k-1} \qquad \Leftrightarrow \qquad \frac{b_g}{d_g} + \cdots + \frac{b_{-k}}{d_{-k}} \;=\; 0.$$

Finally, considering the notation (34) we easily get

$$\int_a^b s_k^{\mathbf{b}}(x)\,\mathrm{d}x \;=\; 0 \qquad \Leftrightarrow \qquad \sum_{i=-k}^{g} b_i\,(\lambda_{i+k+1} - \lambda_i) \;=\; 0.$$

Algorithm for finding a spline with zero-integral
To find an arbitrary spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ with zero-integral

1. Choose $g+k$ arbitrary B-spline coefficients $b_i \in \mathbf{R}$, $i = -k\ldots, j-1, j+1, \ldots, g$,
2. Compute

$$b_j \;=\; \frac{-1}{\lambda_{j+k+1} - \lambda_j} \sum_{\substack{i=-k \\ i\neq j}}^{g} b_i\,(\lambda_{i+k+1} - \lambda_i)\,.$$

It can be easily checked that for these B-spline coefficients the condition

$$\sum_{i=-k}^{g} b_i\,(\lambda_{i+k+1} - \lambda_i) \;=\; 0$$

is fulfilled, and, with respect to Theorem 5.1, the spline $s_k(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x)$ satisfies the condition $\int_a^b s_k(x)\,\mathrm{d}x = 0$.

Proof of Proposition 5.2

Denote by $\mathbf{a}_{(s)}$ the sth row of the matrix product $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $s = 0, \ldots, r$, $d_j = \lambda_j - \lambda_{j-k-1}$, $j = 1, \ldots, g+k+1$, and $\mathbf{1}_{g+k+1}$ a vector of $g+k+1$ ones. Then

$$\begin{aligned}
\sum_{j=1}^{g+k+1} \widehat{\beta}_{js} d_j &= d_1 \mathbf{a}_{(s)}\mathbf{Y}_1 + d_2 \mathbf{a}_{(s)}\mathbf{Y}_2 + \cdots + d_{g+k+1}\mathbf{a}_{(s)}\mathbf{Y}_{g+k+1} = \\
&= \mathbf{a}_{(s)}(d_1\mathbf{Y}_1, d_2\mathbf{Y}_2, \ldots, d_{g+k+1}\mathbf{Y}_{g+k+1})\mathbf{1}_{g+k+1} = \\
&= \mathbf{a}_{(s)}\left(\sum_{j=1}^{g+k+1} Y_{1,j}d_j, \sum_{j=1}^{g+k+1} Y_{2,j}d_j, \ldots, \sum_{j=1}^{g+k+1} Y_{g+k+1,j}d_j\right) = 0.
\end{aligned}$$

Appendix C: smoothing splines for density functions

In this appendix, we briefly describe the computation of B-spline coefficients for a smoothing spline with zero-integral; for more details see Machalová et al. (2016). Assume that the data (x_i, y_i), $a \le x_i \le b$, the weights $w_i \ge 0$, $i = 1, \ldots, n$, $n \ge g + 1$ and the parameter $\alpha \in (0, 1)$ are given. For an arbitrary $l \in \{1, \ldots, k-1\}$ our aim is to find a spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$, which minimizes the functional

$$J_l(s_k) = \alpha \int_a^b \left[s_k^{(l)}(x)\right]^2 \,\mathrm{d}x + \sum_{i=1}^{n} w_i \left[y_i - s_k(x_i)\right]^2 \tag{39}$$

and fulfils the condition

$$\int_a^b s_k(x)\,\mathrm{d}x = 0.$$

In Machalová et al. (2016) it was shown that this spline is given by the formula

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

where the vector of B-spline coefficients $\mathbf{b}^* = (b_{-k}^*, \ldots, b_g^*)'$ is obtained by

$$\mathbf{b}^* = \mathbf{DK}\left[\alpha\,(\mathbf{DK})'\,\mathbf{N}_{kl}\mathbf{DK} + (\mathbf{C}_{k+1}(\mathbf{x})\mathbf{DK})'\,\mathbf{W}\mathbf{C}_{k+1}(\mathbf{x})\mathbf{DK}\right]^+ \mathbf{K}'\mathbf{D}'\mathbf{C}_{k+1}'(\mathbf{x})\mathbf{W}\mathbf{y}.$$

Here, $\mathbf{A}^+$ denotes the Moore-Penrose pseudo-inverse of a matrix $\mathbf{A}$, $\mathbf{W} = \mathrm{diag}(\mathbf{w})$, $\mathbf{w} = (w_1, \ldots, w_n)'$, $\mathbf{y} = (y_1, \ldots, y_n)'$,

$$\mathbf{C}_{k+1}(\mathbf{x}) = \begin{pmatrix} B_{-k}^{k+1}(x_1) & \ldots & B_g^{k+1}(x_1) \\ \vdots & \ddots & \vdots \\ B_{-k}^{k+1}(x_n) & \ldots & B_g^{k+1}(x_n) \end{pmatrix} \in \mathbb{R}^{n,g+k+1}$$

is the collocation matrix,

$$\mathbf{D} = (k+1)\,\mathrm{diag}\left(\frac{1}{\lambda_1 - \lambda_{-k}}, \ldots, \frac{1}{\lambda_{g+k+1} - \lambda_g}\right) \in \mathbb{R}^{g+k+1,g+k+1}$$

and

$$\mathbf{K} = \begin{pmatrix} 1 & 0 & 0 & \cdots & -1 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{g+k+1,g+k+1}.$$

The matrix $\mathbf{N}_{kl} = \mathbf{S}_l'\mathbf{M}_{kl}\mathbf{S}_l$ is positive semidefinite, with

$$\mathbf{M}_{kl} = \begin{pmatrix} \left(B_{-k+l}^{k+1-l}, B_{-k+l}^{k+1-l}\right) & \ldots & \left(B_g^{k+1-l}, B_{-k+l}^{k+1-l}\right) \\ \vdots & & \vdots \\ \left(B_{-k+l}^{k+1-l}, B_g^{k+1-l}\right) & \ldots & \left(B_g^{k+1-l}, B_g^{k+1-l}\right) \end{pmatrix} \in \mathbb{R}^{g+k+1-l,g+k+1-l}.$$

The symbol

$$\left(B_i^{k+1-l}, B_j^{k+1-l}\right) = \int_a^b B_i^{k+1-l}(x) B_j^{k+1-l}(x)\,\mathrm{d}x$$

stands for the scalar product of B-splines in $L^2([a,b])$ space. The matrix $\mathbf{S}_l$ is an upper triangular matrix such that $\mathbf{S}_l = \mathbf{D}_l\mathbf{L}_l \dots \mathbf{D}_1\mathbf{L}_1 \in \mathbb{R}^{g+k+1-l,g+k+1}$, and $\mathbf{D}_j \in \mathbb{R}^{g+k+1-j,g+k+1-j}$ is a diagonal matrix such that

$$\mathbf{D}_j = (k+1-j)\,\mathrm{diag}\,(d_{-k+j},\dots,d_g)$$

with

$$d_i = \frac{1}{\lambda_{i+k+1-j} - \lambda_i} \quad \forall i = -k+j,\dots,g$$

and

$$\mathbf{L}_j := \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{g+k+1-j,g+k+2-j}.$$

Appendix D: Aggregated data

weight[grams]	log(weight)	Proportions of log(C18) classes, $W_{i1},\ldots,W_{i10}, i=1,\ldots,10$									
w_1 2324	7.751	0.008	0.024	0.051	0.110	0.165	0.173	0.194	0.153	0.088	0.035
w_2 2793	7.935	0.010	0.029	0.076	0.130	0.214	0.223	0.165	0.091	0.045	0.017
w_3 2964	7.994	0.014	0.041	0.093	0.145	0.214	0.201	0.149	0.089	0.048	0.008
w_4 3095	8.037	0.015	0.040	0.079	0.152	0.252	0.168	0.152	0.083	0.042	0.015
w_5 3200	8.071	0.004	0.041	0.126	0.169	0.171	0.196	0.171	0.089	0.027	0.008
w_6 3309	8.105	0.008	0.052	0.107	0.147	0.200	0.172	0.182	0.095	0.031	0.006
w_7 3425	8.139	0.012	0.077	0.090	0.171	0.204	0.185	0.144	0.075	0.037	0.006
w_8 3549	8.175	0.025	0.045	0.117	0.187	0.251	0.179	0.096	0.051	0.033	0.016
w_9 3709	8.218	0.021	0.050	0.131	0.182	0.230	0.163	0.129	0.054	0.025	0.015
w_{10} 4103	8.319	0.021	0.054	0.114	0.186	0.269	0.164	0.106	0.046	0.033	0.006
midpoints of intervals t_j		-2.836	-2.636	-2.437	-2.237	-2.037	-1.838	-1.638	-1.439	-1.239	-1.039

Table 2: Proportions of log(C18) classes within 10 log(weight) groups for girls. The values $t_j, j=1,\ldots,10$ are the midpoints of the log(C18) subintervals of $I_b=[-2.936,-0.939]$.

weight[grams]	log(weight)	Proportions of log(C18) classes, $W_{i1},\ldots,W_{i10}, i=1,\ldots,10$									
w_1 2380	7.775	0.002	0.024	0.057	0.124	0.191	0.185	0.179	0.130	0.067	0.039
w_2 2906	7.974	0.021	0.046	0.087	0.137	0.218	0.216	0.133	0.091	0.033	0.019
w_3 3084	8.034	0.017	0.049	0.101	0.181	0.202	0.184	0.142	0.089	0.027	0.008
w_4 3224	8.078	0.014	0.052	0.104	0.176	0.220	0.185	0.149	0.058	0.035	0.008
w_5 3345	8.115	0.017	0.035	0.123	0.175	0.221	0.217	0.121	0.069	0.017	0.006
w_6 3455	8.147	0.027	0.056	0.134	0.165	0.228	0.172	0.130	0.056	0.031	0.002
w_7 3569	8.180	0.023	0.061	0.126	0.178	0.232	0.195	0.103	0.050	0.025	0.006
w_8 3699	8.216	0.019	0.075	0.140	0.190	0.202	0.165	0.131	0.056	0.015	0.008
w_9 3874	8.262	0.023	0.046	0.137	0.174	0.226	0.207	0.112	0.052	0.019	0.004
w_{10} 4232	8.350	0.029	0.079	0.126	0.229	0.231	0.167	0.089	0.029	0.014	0.008
midpoints of intervals t_j		-2.711	-2.506	-2.301	-2.096	-1.891	-1.685	-1.480	-1.275	-1.070	-0.865

Table 3: Proportions of log(C18) classes within 10 log(weight) groups for boys. The values $t_j, j=1,\ldots,10$ are the midpoints of log(C18) subintervals of $I_b=[-2.813,-0.763]$.

weight[grams]	log(weight)	Clr transformation of log(C18) classes, $Z_{i1},\ldots,Z_{i10}, i=1,\ldots,10$									
w_1 2324	7.751	-2.185	-1.086	-0.313	0.454	0.860	0.906	1.024	0.786	0.235	-0.681
w_2 2793	7.935	-1.916	-0.818	0.138	0.679	1.175	1.219	0.917	0.324	-0.390	-1.328
w_3 2964	7.994	-1.585	-0.487	0.340	0.786	1.178	1.113	0.813	0.298	-0.312	-2.145
w_4 3095	8.037	-1.505	-0.540	0.129	0.785	1.290	0.881	0.785	0.176	-0.494	-1.505
w_5 3200	8.071	-2.687	-0.336	0.794	1.086	1.097	1.235	1.097	0.448	-0.741	-1.994
w_6 3309	8.105	-2.059	-0.149	0.562	0.886	1.190	1.044	1.098	0.447	-0.672	-2.346
w_7 3425	8.139	-1.715	0.182	0.343	0.982	1.156	1.057	0.810	0.156	-0.563	-2.409
w_8 3549	8.175	-1.015	-0.445	0.514	0.984	1.279	0.941	0.311	-0.322	-0.747	-1.501
w_9 3709	8.218	-1.186	-0.326	0.635	0.970	1.203	0.859	0.621	-0.252	-1.019	-1.505
w_{10} 4103	8.319	-1.089	-0.154	0.591	1.078	1.448	0.956	0.521	-0.309	-0.653	-2.388
midpoints of intervals t_j		-2.836	-2.636	-2.437	-2.237	-2.037	-1.838	-1.638	-1.439	-1.239	-1.039

Table 4: Clr transformation of log(C18) classes within the 10 log(weight) groups for girls. The values $t_j, j=1,\ldots,10$ are the midpoints of the log(C18) subintervals of $I_g=[-2.936,-0.939]$.

weight[grams]	log(weight)	Clr transformation of log(C18) classes, $Z_{i1}, \ldots, Z_{i10}, i = 1, \ldots, 10$									
w_1 2380	7.775	−3.434	−0.949	−0.066	0.710	1.141	1.110	1.077	0.756	0.093	−0.438
w_2 2906	7.974	−1.233	−0.453	0.176	0.632	1.096	1.087	0.603	0.219	−0.798	−1.329
w_3 3084	8.034	−1.327	−0.305	0.427	1.008	1.120	1.030	0.766	0.304	−0.885	−2.138
w_4 3224	8.078	−1.560	−0.211	0.483	1.004	1.230	1.058	0.837	−0.105	−0.616	−2.120
w_5 3345	8.115	−1.228	−0.535	0.734	1.086	1.320	1.302	0.718	0.158	−1.228	−2.327
w_6 3455	8.147	−0.796	−0.067	0.814	1.020	1.344	1.065	0.785	−0.067	−0.662	−3.435
w_7 3569	8.180	−1.016	−0.035	0.689	1.032	1.295	1.125	0.489	−0.242	−0.935	−2.402
w_8 3699	8.216	−1.199	0.162	0.789	1.094	1.153	0.953	0.718	−0.134	−1.422	−2.115
w_9 3874	8.262	−0.936	−0.243	0.841	1.079	1.341	1.252	0.639	−0.125	−1.119	−2.728
w_{10} 4232	8.350	−0.739	0.267	0.727	1.324	1.332	1.007	0.382	−0.739	−1.501	−2.061
midpoints of intervals t_j		−2.711	−2.506	−2.301	−2.096	−1.891	−1.685	−1.480	−1.275	−1.070	−0.865

Table 5: Clr transformation of the log(C18) classes within the 10 log(weight) groups for boys. The values $t_j, j = 1, \ldots, 10$ are the midpoints of the log(C18) subintervals of $I_b = [-2.813, -0.763]$.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London.

van den Boogaart, K., Egozcue, J., Pawlowsky-Glahn, V., 2010. Bayes linear spaces. Statistics and Operations Research Transactions 34, 201–222.

van den Boogaart, K., Egozcue, J., Pawlowsky-Glahn, V., 2014. Bayes Hilbert spaces. Australian & New Zealand Journal of Statistics 54, 171–194. doi:10.1111/anzs.12074.

de Boor, C., 1978. A Practical Guide to Splines. Springer, New York.

Delicado, P., 2011. Dimensionality reduction when data are density functions. Computational Statistics and Data Analysis 55, 401–420. doi:10.1016/j.csda.2010.05.008.

Dierckx, P., 1993. Curve and surface fitting with splines. Clarendon Press, Oxford .

Egozcue, J., Díaz-Barrero, J., Pawlowsky-Glahn, V., 2006. Hilbert space of probability density functions based on Aitchison geometry. Acta Mathematica Sinica, English Series 22, 1175–1182. doi:10.1007/s10114-005-0678-2.

Egozcue, J., i Estadella, J.D., Pawlowsky-Glahn, V., Hron, K., Filzmoser, P., 2012. Simplicial regression. The normal model. Journal of Applied Probability and Statistics 6, 87–108.

Faraway, J., 1997. Regression analysis for a functional response. Technometrics 3, 254–261. doi:10.2307/1271130.

Fišerová, E., Kubáček, L., Kunderová, P., 2007. Linear Statistical Models: Regularity and Singularities. Academia, Prague.

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P., 2016. Simplicial principal component analysis for density functions in bayes spaces. Computational Statistics and Data Analysis 94, 330–350. doi:10.1016/j.csda.2015.07.007.

Johnson, R., Wichern, D., 2007. Applied Multivariate Statistical Analysis (6th edn). Prentice-Hall, London.

Machalová, J., Hron, K., Monti, G., 2016. Preprocessing of centred logratio transformed density functions using smoothing splines. Journal of Applied Statistics 43, 1419–1435. doi:10.1080/02664763.2015.1103706.

Martín-Fernández, J., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J., 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical Modelling 15, 134–158.

Menafoglio, A., Guadagnini, A., Secchi, P., 2014. A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. Stochastic Environmental Research and Risk Assessment 28, 1835–1851. doi:10.1007/s00477-014-0849-8.

Menafoglio, A., Guadagnini, A., Secchi, P., 2016a. Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a bayes space approach. Water Resources Research 52, 5708–5726. doi:10.1002/2015WR018369.

Menafoglio, A., Secchi, P., Guadagnini, A., 2016b. A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. Mathematical Geosciences 48, 463–485. doi:10.1007/s11004-015-9625-7.

Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. Wiley, Chichester.

Ramsay, J., Silverman, B., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer, New York.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis, 2nd edition. Springer, New York.

Shena, Q., Xub, H., 2007. Diagnostics for linear models with functional responses. Technometrics 1, 26–33.

Sturges, H.A., 1926. The choice of a class interval. Journal of the American Statistical Association 21,153 65–66.