MOX-Report No. 21/2017

Compositional regression with functional response

Talska, R.; Menafoglio, A.; Machalova, J.; Hron, K.; Fiserova, E.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Compositional regression with functional response

R. Talská[1], A. Menafoglio[2*], J. Machalová[1], K. Hron[1] and E. Fišerová[1]

[1]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Olomouc, Czech Republic.
[2]MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy
*alessandra.menafoglio@polimi.it

Abstract

This work addresses the problem of performing functional linear regression when the response variable is represented as a probability density function (PDF). PDFs are interpreted as functional compositions, that 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 feature 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 employed for actual computations. For this purpose, a new key result is proved to characterize B-spline representations in Bayes spaces. We show the potential of the methodological developments on a real case study, dealing with metabolomics data. Here, a bootstrap-based study is also 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). In general, the PDFs are Borel measurable functions that are constrained to be non-negative and to integrate to unity. Nevertheless, 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. In

fact, when changing the value the PDF integrates to a general real constant c (i.e., the measure of the whole), the *relative* information carried by PDFs is preserved – we refer to *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 turns out 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 to a scaled PDF, that 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 only 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 already studied in detail in the seminal book of functional data analysis (Ramsay and Silverman, 2005), the case of functional regression with a distributional response variable has not been systematically elaborated yet. In fact, most techniques developed in that setting to deal with PDFs (e.g., Ramsay and Silverman (2005), Section 6.6) aim to remove the constant-integral constraint, rather than taking into account the key properties of PDFs for further statistical processing. Instead, the key point of our approach is to consider PDFs as elements of a Bayes space, and accordingly work with the geometry of the latter space. 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).

We employ the B-spline representation of clr-transformed data proposed in Machalová et al. (2016) to express discretely observed PDFs as smooth functions. Such representation allows to properly incorporate the zero-integral constraint resulting from the clr-transformed PDFs (van den Boogaart et al., 2014) in the B-spline basis: we here show that such zero-integral constraint is equivalent to a linear constraint on the B-spline coefficients, which needs to be properly taken into account in the further statistical processing. On these bases, we shall introduce (penalized) least squares estimators for the functional regression coefficients and their variability in the Bayes space. We illustrate the potential of our approach through its application to a real study.

The remaining part of the paper is organized as follows. In Section 2 the proba-

bility density functions from perspective of Bayes spaces are introduced together with the so called centred log-ratio transformation which enables one to apply well-known techniques of functional data analysis. The functional linear regression model with functional response and scalar regressors for functional data in L^2 is recalled in Section 3 and its counterpart in Bayes space is proposed in Section 4. Section 5 deals with B-spline representation of PDFs in Bayes spaces and Section 6 with multivariate regression modeling of B-spline coefficients, some specific features are derived in Section 7. In Section 8, the methodological developments are illustrated with real data containing measurements of metabolite concentrations.

2 Probability densities as elements of Bayes spaces

Similarly 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), processing PDFs within the usual L^2 space may lead to meaningless results. Instead, the specific features of densities can be captured through the Bayes space methodology that relies 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, that contributions of Borel sets to the whole mass measure do not change.

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, i.e.,

$$\mathcal{B}^2(I) = \{f : I \to (0, +\infty), \int_I [\log f(t)]^2 dt < \infty\}.$$

To avoid highly technical constructions, we limit to consider compact support $I = [a, b] \subset \mathbf{R}$. For a density f, we denote by $\mathcal{C}(f)$ the unit-integral representative within its equivalence class in $\mathcal{B}^2(I)$. The operation $\mathcal{C}(f)$ is called *closure*, which result is simply obtained by dividing f by its integral over the interval I.

Further, we introduce operations in $\mathcal{B}^2(I)$, called *perturbation* and *powering*, which play the role of sum and multiplication by a scalar. Consider two densities $f, g \in \mathcal{B}^2(I)$ and a real number $c \in \mathbf{R}$, perturbation and powering are defined as

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

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. It can be shown (Egozcue et al., 2006; van den Boogaart et al., 2014) that the triple $\left(\mathcal{B}^2(I), \oplus, \odot\right)$ forms a vector space. The Bayes inner product is 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

$$\|f\|_{\mathcal{B}} = \sqrt{\langle f, f\rangle_{\mathcal{B}}} = \left[\frac{1}{2\eta}\int_a^b \int_a^b \ln^2 \frac{f(t)}{f(s)} dt\, ds\right]^{\frac{1}{2}}$$

and

$$d_{\mathcal{B}}(f, g) = \|f \ominus g\|_{\mathcal{B}} = \left[\frac{1}{2\eta}\int_a^b \int_a^b \left(\ln \frac{f(t)}{f(s)} - \ln \frac{g(t)}{g(s)}\right)^2 dt\, ds\right]^{\frac{1}{2}},$$

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.

In Egozcue et al. (2006) and van den Boogaart et al. (2014) it was shown that the Bayes space $\left(\mathcal{B}^2(I), \oplus, \odot, \langle\cdot,\cdot\rangle_{\mathcal{B}}\right)$ forms a separable Hilbert space. Accordingly, 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*. The clr transformation of a PDF $f \in \mathcal{B}^2(I)$ is the real-valued function $f_c \in L^2(I)$ defined as

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

The clr representation 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. The original density $f \in \mathcal{B}^2(I)$ can be obtained from the corresponding clr transform $f_c \in L^2(I)$ through the inverse transformation

$$f(t) = \operatorname{clr}^{-1}[f_c](t) = \mathcal{C}(\exp[f_c])(t), \quad t \in I. \tag{2}$$

Finally, we point out that the following important properties of the isometric isomorphism (1) hold

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

and

$$\langle f, g\rangle_{\mathcal{B}} = \langle f_c, g_c\rangle_2, \quad \|f\|_{\mathcal{B}} = \|f_c\|_2, \quad d_{\mathcal{B}}(f, g) = d_2(f_c, g_c),$$

where $\langle\cdot,\cdot\rangle_2$, $\|\cdot\|_2$ and $d_2(\cdot,\cdot)$ denote inner product, norm and distance in $L^2(I)$ respectively. Intuitively, 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

In real studies the functional variables of interest are frequently driven by one or more covariates, either of real or of functional nature. A large body of literature has been

developed on both theoretical and applied issues related to functional linear models (Faraway, 1997; Shena and Xub, 2007); several approaches to linear regression with functional response and multivariate covariates are broadly discussed in Ramsay and Silverman (2005). We here review the key notions on functional linear models with scalar regressors that we deem useful for our developments, by following Ramsay and Silverman (2005, Chapter 13), to which the reader is 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, i.e., $x_0 = 1$. Let us 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 is made of ones if the intercept is included) and a p-dimensional vector of unknown functional regression parameters $\boldsymbol{\beta}(t)$ in $L^2(I)$, which is unknown and has to be estimated. Furthermore, let $\boldsymbol{\varepsilon}(t)$ be an N-dimensional vector of i.i.d. (functional) random errors with zero-mean in $L^2(I)$. The functional linear model is expressed as

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{r} x_{ij}\beta_j(t) + \varepsilon_i(t), \quad i = 1, \ldots, N, \tag{3}$$

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,

$$\text{SSE}(\boldsymbol{\beta}) = \int_I \left[\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)\right]' \left[\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)\right] dt. \tag{4}$$

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

$$\text{PENSSE}(\boldsymbol{\beta}) = \int_I \left[\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)\right]' \left[\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)\right] dt + \lambda \int_I \left[L\boldsymbol{\beta}(s)\right]' \left[L\boldsymbol{\beta}(s)\right] ds, \tag{5}$$

with L a linear differential operator and λ a smoothing parameter. Setting a low value of λ leads to a better fit to the observed data at the expense of a higher roughness of the estimates. Conversely, for higher values of λ a worse fit is obtained, but the smoothness of functions $\widehat{\beta}_j(t)$ is increased.

Several computational methods have been proposed in the literature to minimize (4) or (5). 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. Suppose that $y_i(t)$ and $\beta_j(t)$ admit the representations

$$y_i(t) = \sum_{k}^{K_y} c_{ik}\varphi_k(t), \quad \beta_j(t) = \sum_{k}^{K_\beta} b_{jk}\psi_k(t), \tag{6}$$

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

$$\begin{aligned}\text{PENSSE}(\boldsymbol{\beta}) = &\int_I \left[\mathbf{C}\boldsymbol{\varphi}(t) - \mathbf{X}\mathbf{B}\boldsymbol{\varphi}(t)\right]' \left[\mathbf{C}\boldsymbol{\varphi}(t) - \mathbf{X}\mathbf{B}\boldsymbol{\varphi}(t)\right] dt \\ &+\lambda \int_I \left[L\mathbf{B}\boldsymbol{\varphi}(s)\right]' \left[L\mathbf{B}\boldsymbol{\varphi}(s)\right] ds. \end{aligned} \tag{7}$$

Note that setting $\lambda = 0$ yields the reformulation of (4) 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 (7) with respect to $\mathbf{B}$ it can be shown that the estimation of $\mathbf{B}$ is found as solution of the linear system

$$\left(\mathbf{X}'\mathbf{X}\mathbf{B}\mathbf{Q} + \lambda\mathbf{B}\mathbf{P}\right) = \mathbf{X}'\mathbf{C}\mathbf{Q}. \tag{8}$$

Note that in this setting, the same level of smoothness is imposed for all the $\beta_j(t)$, $j = 0, \ldots, r$. System (8) can be equivalently reformulated using the Kronecker product $\otimes$ as

$$\left[\mathbf{Q} \otimes \left(\mathbf{X}'\mathbf{X}\right) + \mathbf{P} \otimes \lambda\mathbf{I}\right] \text{vec}(\mathbf{B}) = \text{vec}\left(\mathbf{X}'\mathbf{C}\mathbf{Q}\right). \tag{9}$$

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

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 model (3). 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

$$y_i(t) = \beta_0(t) \oplus \bigoplus_{j=1}^{r} \left[x_{ij} \odot \beta_j\right](t) \oplus \varepsilon_i(t) \tag{10}$$

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 $\mathcal{B}^2(I)$

$$\text{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\varepsilon_i\|_{\mathcal{B}}^2 = \sum_{i=1}^{N} \left\| \bigoplus_{j=0}^{r} \left[x_{ij} \odot \beta_j\right] \ominus y_i \right\|_{\mathcal{B}}^2. \tag{11}$$

Note that (11) is the counterpart of SSE (4) in the Bayes Hilbert space; in fact, it also represents the analogue of compositional SSE (Egozcue et al., 2012) in infinite dimensions. Applying clr transformation (1) to both sides of the model (11) yields

$$\mathrm{clr}(y_i)(t) = \mathrm{clr}(\beta_0)(t) + \sum_{j=1}^{r} \left[x_{ij} \cdot \mathrm{clr}(\beta_j)\right](t) + \mathrm{clr}(\varepsilon_i)(t), \quad i = 1, \ldots, N, \quad (12)$$

that enables one to reformulate the objective SSE (11) equivalently in the L^2 sense as

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\mathrm{clr}(\varepsilon_i)\|_2^2 = \sum_{i=1}^{N} \left\| \sum_{j=0}^{r} [x_{ij} \cdot \mathrm{clr}(\beta_j)] - \mathrm{clr}(y_i) \right\|_2^2. \quad (13)$$

In this work, we focus 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 we shall discuss in Section 7. We note that alternatively one could develop PENSSE by closely follow the arguments here presented.

Note that 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 (13) need to follow the zero-integral constraint, i.e.,

$$\int_I \mathrm{clr}(y_i(t))\, dt = 0; \quad \int_I \mathrm{clr}(\beta_j(t))\, dt = 0. \quad (14)$$

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. 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

$$\mathrm{clr}(y_i(t)) = \sum_{k}^{K} c_{ik}\varphi_k(t); \quad \mathrm{clr}(\beta_j(t)) = \sum_{k}^{K} b_{jk}\varphi_k(t) \quad (15)$$

or, in matrix notation, $\mathrm{clr}(y_i(t)) = \mathbf{c}_i'\boldsymbol{\varphi}(t)$ and $\mathrm{clr}(\beta_j(t)) = \mathbf{b}_j'\boldsymbol{\varphi}(t)$. In this case, the zero integral constraints in (14) reads

$$\int_I \mathrm{clr}(y_i(t))\, dt = \int_I \sum_{k}^{K} c_{ik}\varphi_k(t)\, dt = 0; \quad \int_I \mathrm{clr}(\beta_j(t))\, dt = \int_I \sum_{k}^{K} b_{jk}\varphi_k(t)\, dt = 0. \quad (16)$$

These constraints need to be carefully taken into account when estimating the linear model (10), as they may turn in linear constraints on the coefficients $\{c_{ik}\}, \{b_{jk}\}$ and consequently on model singularities. We discuss this and its implications in the next Sections, in the light of the key result proved in Section 5.

5 The B-spline representation for density functions in Bayes spaces

The clr transformation of both the response PDFs and the regression coefficients in model (12) need to fulfill the zero-integral constraint. As shown in Machalová et al.

(2016), it is possible to find an explicit expression for the B-spline representation of a clr-transformed PDF, fulfilling the zero-integral constraint. In this section we recall the basic notions on smoothing B-splines and show that the zero integral constraint on clr induces a corresponding constraint on B-spline coefficients, that in turn is characterizing this class of B-splines. This fact will be useful to reduce the dimensionality of the B-spline representation of densities proposed in Machalová et al. (2016) without loss of information and thus avoid singularity of the resulting regression model.

Studying B-spline representations for density data is key from the application viewpoint, since in most practical situations the PDFs under study are sampled in terms of histogram data. Indeed, for each of the (theoretical) 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 $y_i, i = 1,...,n$. In fact, 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 an unconstrained Euclidean space, one may employ the discrete version of the clr transformation (1), that is (Aitchison, 1986)

$$Z_{ij} = \ln \frac{W_{ij}}{\sqrt[D]{\prod_{j=1}^{D} W_{ij}}}, \quad j = 1,\ldots,D. \tag{17}$$

Denote by $\underline{\mathbf{Z}} = (Z_{ij})$ the matrix of clr-transformed raw data. Similarly as in FDA, as a first step of the analysis one may want to smooth the observations, to obtain an estimate of the underlying continuous density from raw data. For this purpose, following Machalová et al. (2016), we here consider smoothing splines.

First, let us recall the basic notions on B-splines that we deem useful for the following developments (see de Boor (1978), Dierckx (1993) for details). Let the sequence of knots

$$\Delta\lambda := \lambda_0 = a < \lambda_1 < \ldots < \lambda_g < b = \lambda_{g+1}$$

be given. In the following, $\mathcal{S}_k^{\Delta\lambda}[a,b]$ denotes the vector space of polynomial splines of degree $k > 0$, defined on a finite interval $I = [a,b]$ with the sequence of knots $\Delta\lambda$. It is known that $dim\left(\mathcal{S}_k^{\Delta\lambda}[a,b]\right) = g + k + 1$. Then every spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ has a unique representation as

$$s_k(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x). \tag{18}$$

For this representation it is necessary to add some additional knots to be able to construct all basis functions of $\mathcal{S}_k^{\Delta\lambda}[a,b]$. Without loss of generality, we can add knots

$$\lambda_{-k} = \cdots = \lambda_{-1} = \lambda_0, \qquad \lambda_{g+1} = \lambda_{g+2} = \cdots = \lambda_{g+k+1}. \tag{19}$$

Vector $\mathbf{b} = (b_{-k},\ldots,b_g)'$ is called the *vector of B-spline coefficients* of $s_k(x)$, functions $B_i^{k+1}(x)$, $i = -k,\ldots,g$, are *B-splines of degree* k and form basis in $\mathcal{S}_k^{\Delta\lambda}[a,b]$.

In line with previous considerations, to deal with density data through clr transforms, one needs to build the spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$, in a way that guarantees that it has zero integral on the finite interval $[a,b]$. That is, for $\mathbf{Z}_{(i)} = (Z_{i1},\ldots,Z_{iD})'$, with $i = 1,\ldots,n$, one needs to look for the smoothing spline which satisfies the condition

$$\int_a^b s_k^i(x)\,\mathrm{d}x = 0, \tag{20}$$

and best approximates the data, according to an appropriate criterion (see Appendix B for further details). As proved in Machalová et al. (2016), the *optimal* smoothing spline admits a unique representation

$$s_k^i(x) = \sum_{j=-k}^{g} Y_{i,j+k+1} B_j^{k+1}(x), \tag{21}$$

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

$$\mathbf{Y}_{(i)} = \mathbf{V}\mathbf{Z}_{(i)}, \quad i = 1,\ldots,n. \tag{22}$$

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, if a penalized criterion is chosen (Machalová et al., 2016). If the same B-spline basis system is used for all the data, (22) can be expressed in matrix notation as

$$\underline{\mathbf{Y}} \;=\; \underline{\mathbf{Z}}\mathbf{V}', \tag{23}$$

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

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

The following Theorem 5.1 states a necessary and sufficient condition for a vector $\mathbf{b}$ (e.g., a candidate for $\mathbf{Y}_{(i)}$) to be a vector of B-spline coefficients for a spline with zero integral.

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 A. In the light of Theorem 5.1, it is easy to see that vector $\mathbf{b}$ is orthogonal to the vector $\mathbf{n} = (\lambda_1 - \lambda_{-k},\ldots,\lambda_{g+k+1} - \lambda_g)'$, that 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

$$\sum_{j=1}^{g+k+1} Y_{ij}(\lambda_j - \lambda_{j-k-1}) = 0. \tag{24}$$

Whenever the same B-spline basis is employed for all the data – as it is usually the case – the linear constraint (24) turns into a model singularity, as we shall show in the next Section.

6 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 model (10) in the form of a multivariate regression model (Johnson and Wichern, 2007). 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 (10) the multivariate linear regression model of the form

$$\underline{\mathbf{Y}}_{(n\times(g+k+1))} = \mathbf{X}_{(n\times p)}\mathbf{B}_{(p\times(g+k+1))} + \underline{\boldsymbol{\varepsilon}}_{(n\times(g+k+1))} \tag{25}$$

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.,

$$\mathrm{cov}(\mathbf{Y}_{(i)}, \mathbf{Y}_{(j)}) = \mathbf{0}_{((g+k+1)\times(g+k+1))},\ i \neq j, \quad \mathrm{var}(\mathbf{Y}_{(i)}) = \boldsymbol{\Sigma}_{((g+k+1)\times(g+k+1))},$$

for $i = 1, \ldots N$.

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

$$\widehat{\mathbf{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{g+k+1}), \tag{26}$$

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 that implies an alternative estimation of columns of $\mathbf{B}$ as

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

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

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

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 theorem.

Theorem 6.1 *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* $s = 0, \ldots, r$.

Note that this constraint introduces a singularity into regression model (25). To avoid the necessity of considering singular regression models (Fišerová et al., 2007), orthonormal coordinates from the B-spline coefficients may be considered, in the light of the results of Section 5. The vectors $\mathbf{Y}_{(i)}$, $i = 1, ..., n$, form a hyperplane $\mathcal{H}$ of dimension $g + k$, orthogonal to the normal vector

$$\mathbf{n} = (\lambda_1 - \lambda_{-k}, \ldots, \lambda_{g+k+1} - \lambda_g)',$$

that only depends on the knots positions. For $\mathcal{H}$ one may build an orthonormal basis, express $\mathbf{Y}_{(i)}$, $i = 1, ..., n$, through the coordinates of such a basis – thus removing the singularity due to the linear constraints induced by (14) – and then use the regularized representation for the purpose of further computations. 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, that in turn correspond to the Simplicial Functional Principal Components of the smoothed densities $y_1(t), ..., y_n(t)$ (Hron et al., 2016).

7 Smoothing splines and regression: the relation with the multivariate setting

A natural question which may arise 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, and (b) a compositional regression (Egozcue et al., 2012) is applied, estimating the model

$$\mathbf{Z}_i \;=\; \boldsymbol{\beta}_0^{(Z)} + \sum_{j=1}^{r} \boldsymbol{\beta}_j^{(Z)} x_{ij} + \boldsymbol{\epsilon}_i, \tag{27}$$

and the estimates (or predictions) of $\underline{\mathbf{Z}}$ are smoothed afterward. In particular, we here show that the following scheme holds true

$$\begin{array}{ccc} \mathbf{Z} & \xrightarrow{smoothing} & \underline{\mathbf{Y}} \\ {\scriptstyle regression}\downarrow & & \downarrow{\scriptstyle regression} \\ \widehat{\mathbf{Z}} & \xrightarrow[smoothing]{} & \widehat{\underline{\mathbf{Y}}} \end{array} \tag{28}$$

Recall that, from (26), the matrix of predicted coefficients $\underline{\mathbf{Y}}$ is obtained as

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

Similarly, for model (27) one has

$$\widehat{\underline{\mathbf{Z}}} \;=\; \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Z}}. \tag{29}$$

Plugging-in (23) in (29) 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 predicted data $\widehat{\mathbf{Z}}_{\mathbf{i}}$, $i = 1, \ldots, n$, are considered, the matrix of the corresponding B-spline coefficients is obtained as

$$\widehat{\widehat{\underline{\mathbf{Y}}}} \;=\; \widehat{\underline{\mathbf{Z}}}\mathbf{V}'_Z. \tag{30}$$

In order to guarantee that $\mathbf{V}_Z$ coincides with the matrix $\mathbf{V}$ in (23), one needs just 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 (29), the matrix $\widehat{\widehat{\underline{\mathbf{Y}}}}$ can be written in the form

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

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.

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. The dataset we consider collects the

data about 10000 newborns with standard weights (all the data were anonymised prior to analysis). 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 were 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 $I_b = [-2.813, -0.763]$. On these bases, the vectors of proportions were transformed by using the discrete version of the clr transformation (17).

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. 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., $k = 2$, $l = 1$) with equally spaced sequence of 5 knots $\Delta\lambda_g := [-2.836, -2.387, -1.938, -1.488, -1.039]$, $\Delta\lambda_b := [-2.711, -2.249, -1.788, -1.326, -0.865]$ for girls and boys, respectively. 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 parameter α was set to α = 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 resulting smoothed clr-densities $y_{ci}(t) \in L^2(I)$, $i = 1, ..., 10$, are displayed in Figure 1 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., $y_i(t) = \mathrm{clr}^{-1}\left[y_{ci}\right](t) = \mathcal{C}\left[\exp(y_{ci})\right](t), i = 1, \ldots, 10, t \in I$.

The functional regression model was then built by resorting to separate models for girls and boys, as the supports of the log(C18) distribution differ between the two populations. Thus, for each of the two groups, we separately modeled the dependence

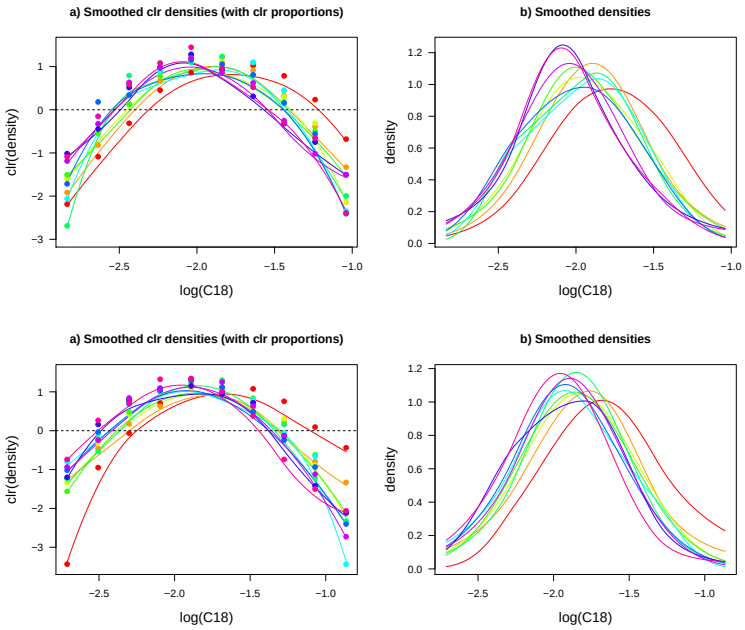

Figure 1: Clr densities and their inverse (i.e., the densities) of log(C18). Girls a)-b), boys c)-d).

of the log(C18) distributions on log(weight) through following linear model in $\mathcal{B}^2(I)$,

$$y_i(t) = \beta_0(t) \oplus [\log(w_i) \odot \beta_1](t) \oplus \varepsilon_i(t), \quad i = 1, \ldots, 10, \tag{31}$$

which is written in $L^2(I)$ as

$$\operatorname{clr}[y_i(t)] = \operatorname{clr}[\beta_0(t)] + \log(w_i) \cdot \operatorname{clr}[\beta_1(t)] + \operatorname{clr}[\varepsilon_i(t)], \quad i = 1, \ldots, 10, \quad t \in I. \tag{32}$$

By considering the same B-spline basis functions $B^3_{-2}(t), \ldots, B^3_3(t)$ for the response $\operatorname{clr}(y(t))$, the regression parameters $\operatorname{clr}[\beta_0(t)]$, $\operatorname{clr}[\beta_1(t)]$ and the error $\operatorname{clr}[\varepsilon(t)]$, model (32) can be written as a multivariate model for the B-spline coefficients $Y_{i1}, \ldots, Y_{i6}$

$$\begin{pmatrix} Y_{1,1} & Y_{1,2} & \cdots & Y_{1,6} \\ Y_{2,1} & Y_{2,2} & \cdots & Y_{2,6} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{10,1} & Y_{10,2} & \cdots & Y_{10,6} \end{pmatrix} = \begin{pmatrix} 1 & \log(w_1) \\ 1 & \log(w_2) \\ \vdots & \vdots \\ 1 & \log(w_{10}) \end{pmatrix} \cdot \begin{pmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{06} \\ \beta_{11} & \beta_{16} & \cdots & \beta_{16} \end{pmatrix}$$
$$+ \begin{pmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,6} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,6} \\ \vdots & \vdots & \ddots & \vdots \\ e_{10,1} & e_{10,2} & \cdots & e_{10,6} \end{pmatrix}, \tag{33}$$

	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	1.145
$\widehat{\sigma}$	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 $\widehat{\sigma} = \left\{ \widehat{\mathrm{var}}(\mathrm{vec}(\widehat{\mathbf{B}})) \right\}_{k,k}$ (grey rows).

or, 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 $\mathrm{clr}[\beta_0(t)]$ and $\mathrm{clr}[\beta_1(t)]$ are displayed in Figure 2, 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 2 (red curves). We first note that the intercept $\beta_0(t)$ is hardly interpretable, as it estimates the expected value of the density of log(C18) when the weight of 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 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 3, 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 distribution of the metabolite seems to be even more dramatic. This indicates a more serious impact of the underweight to predisposition of the metabolic disease for boys.

The fitted curves corresponding to the $N = 10$ observed distributions are displayed in Figure 4 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 determi-

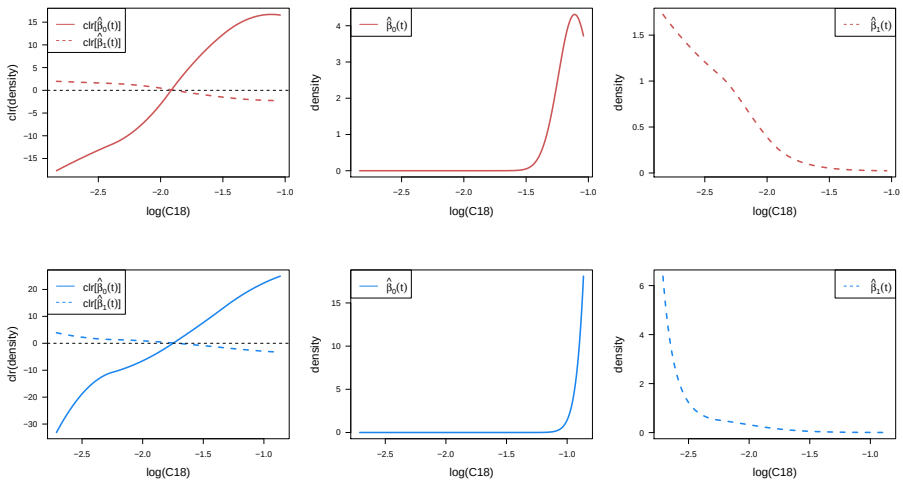

Figure 2: Estimates of regression coefficients. The upper three plots represent the results for the girls' group (red): first, clr estimates of $\beta_0(t)$ and $\beta_1(t)$ in $L^2(I_g)$; second, estimate of $\beta_0(t)$ in $\mathcal{B}^2(I_g)$; third: estimate of $\beta_1(t)$ in $\mathcal{B}^2(I_g)$. Lower three plots represent the results for the boys' group (blue): first, clr estimates of $\beta_0(t)$ and $\beta_1(t)$ in $L^2(I_b)$; second, estimate of $\beta_0(t)$ in $\mathcal{B}^2(I_b)$; third: estimate of $\beta_1(t)$ in $\mathcal{B}^2(I_b)$.

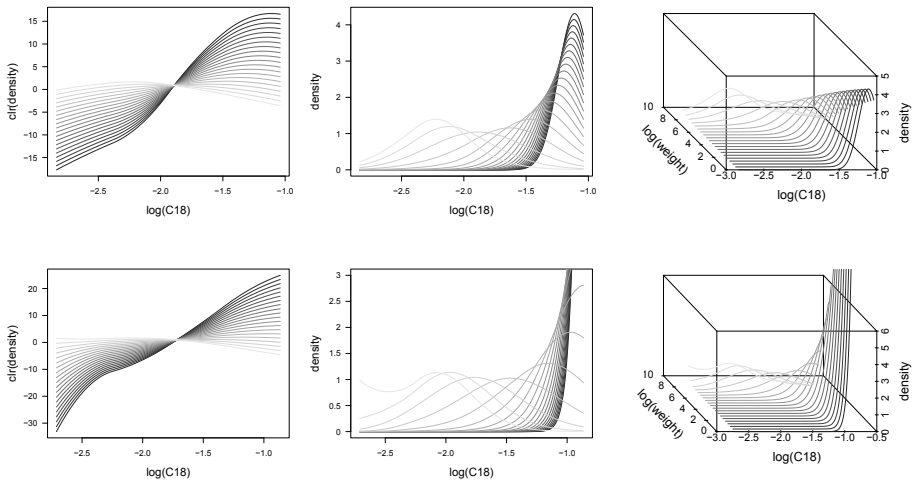

Figure 3: 2D and 3D graphs of predicted distributions for increasing sequence of 20 values of log weights.

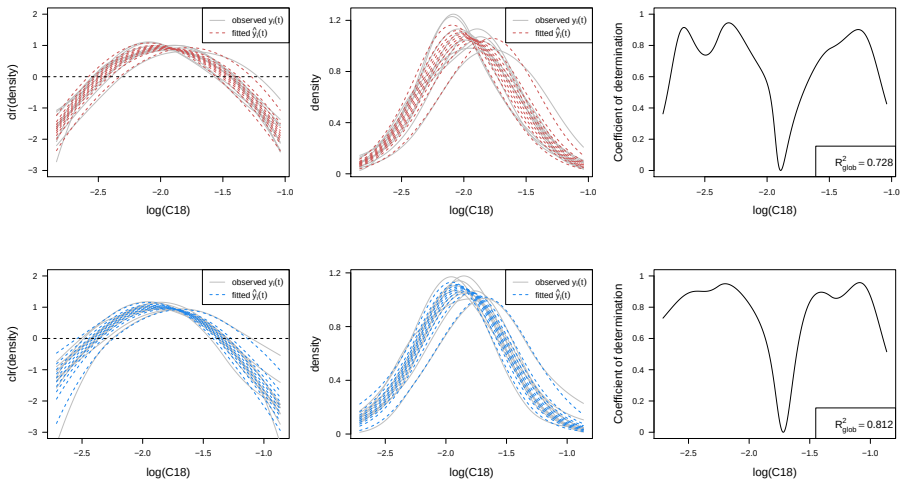

Figure 4: Comparison of observed y (grey) and fitted $\hat{y}$ (girls – red, boys – blue) distributions in L^2 and $\mathcal{B}^2$ (upper and bottom first two figures). Pointwise coefficient of determination (right figures, upper for girls and bottom for boys).

nation $R^2(t)$, $t \in I$, 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 computed 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 4. Although the graphs of 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 bootstrap sample

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

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 5 and 6.

We then used these bootstrap outputs $(\widehat{\beta}_{0r}^{boot}, \widehat{\beta}_{1r}^{boot})_{r=1,\ldots,R}$ to quantify the uncertainty in the fitted model for fixed value of $\log(w)$. Here, two values of weights were chosen to compute 200 fitted curves by using the estimates obtained by bootstrap procedure. The results are displayed in Figure 5 and 6. In both cases, interesting patterns appear by observing the figures. Indeed, most of the uncertainty in β_0 is shown in the right part of domain, whereas for β_1 it is mostly present in the left part of domain. For the girls' case, the bottom two panels of Figure 5 indicate poor fitting for observed distribution corresponding to $\log(w_5)$ which can be also read from pointwise evaluated coefficient of determination (see Figure 4). This can indicate that response might depend on other regressors, not available in this study.

9 Conclusions

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

For the actual estimation of the regression coefficients, we proposed an approach based on a B-spline expansion, properly adapted to deal with density data. Here, we proved a key result on the characterization of the B-spline expansion of clr-transformed data, that 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, that 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.

More in general, the possibility of obtaining estimates of an entire distribution has a great potential from the application viewpoint: our approach enables one to model not only the relation of the mean and the variance of the response on the regressors, but all the moments jointly. Nevertheless, still critical appears the data pre-processing, which requires estimation and smoothing of the response distribution and may introduce additional uncertainty. Ways to account for the latter uncertainty in the estimation procedure are currently under investigation.

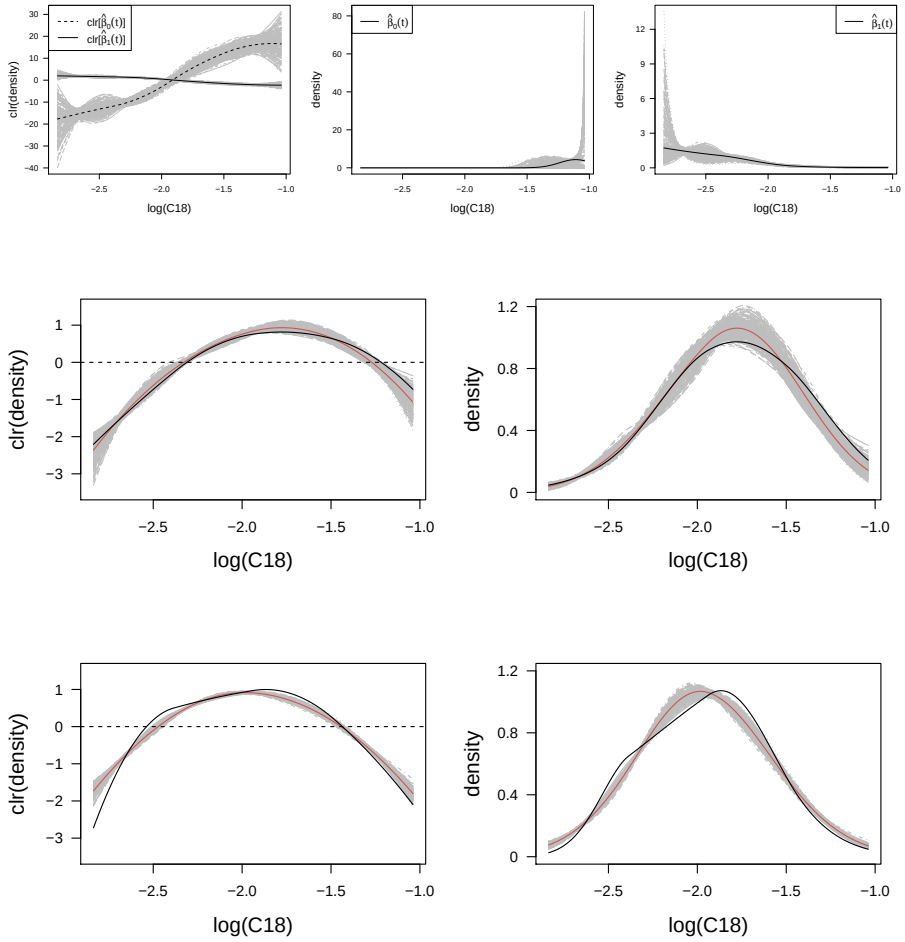

Figure 5: Bootstrap results for the girls' group. Upper three panels: black curves indicate estimates of regression parameters, grey lines indicate the $R = 200$ bootstrap estimates for both the regression parameters. Bottom four panels: black curves indicate observed distributions for w_1 (upper panels) and w_5 (bottom panels), red curves indicate the fitted distribution for w_1 and w_5 by model (31), grey lines indicate the corresponding fitted distributions obtained by bootstrap procedure.

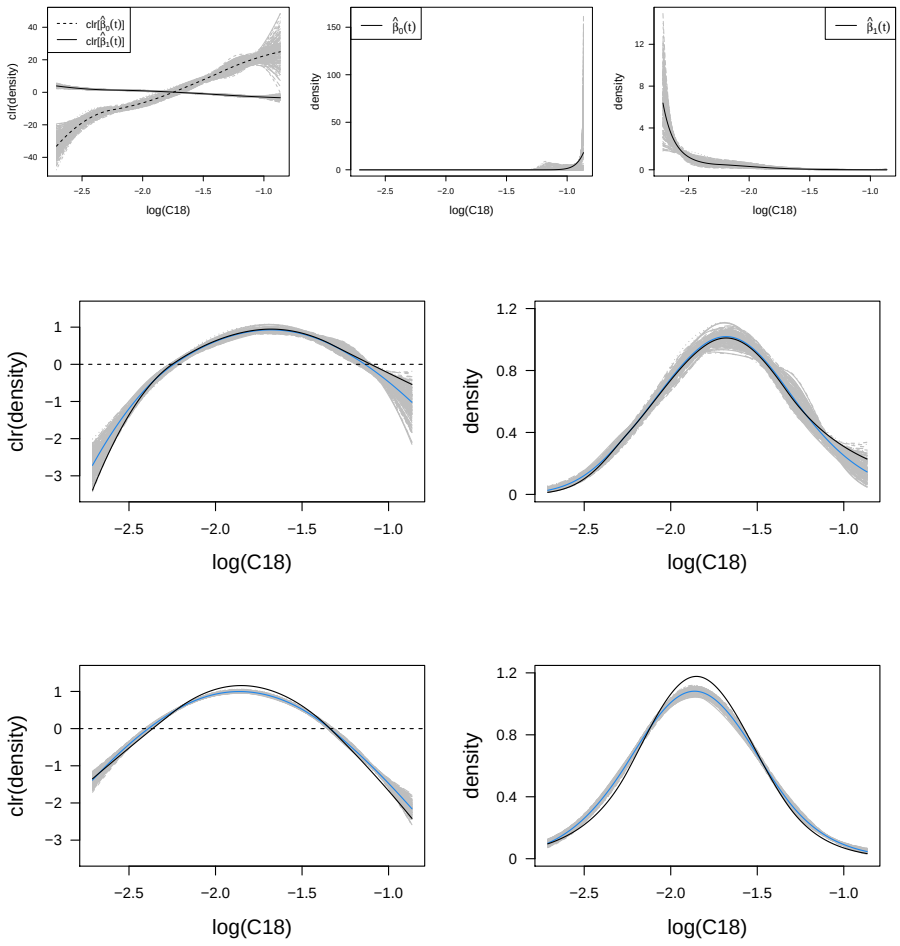

Figure 6: Bootstrap results for the boys' group. Upper three panels: black curves indicate estimates of regression parameters, grey lines indicate the $R = 200$ bootstrap estimates for both the regression parameters. Bottom four panels: black curves indicate observed distributions for w_1 (upper panels) and w_5 (bottom panels), blue curves indicate the fitted distribution for w_1 and w_5 by model (31), grey lines indicate the corresponding fitted distributions obtained by bootstrap procedure.

As a way to assess the estimators uncertainty, we considered a bootstrap resampling method. 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.

Appendix A: 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 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 $\mathbf{c}$, that is

$$s_k^{\mathbf{b}}(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x) \;=\; \frac{\mathrm{d}}{\mathrm{d}x} \sum_{i=-k-1}^{g} c_i B_i^{k+2}(x) \;=\; \frac{\mathrm{d}}{\mathrm{d}x} s_{k+1}^{\mathbf{c}}(x). \tag{34}$$

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

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

To simplify the notation we set

$$d_i = \frac{k+1}{\lambda_{i+k+1} - \lambda_i}, \quad i = -k, \ldots, g; \tag{35}$$

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

$$c_g \;=\; \frac{b_g}{d_g} + \cdots + \frac{b_{-k}}{d_{-k}} + c_{-k-1}. \tag{36}$$

With respect to (34) it is evident that

$$\int_a^b s_k^{\mathbf{b}}(x)\,\mathrm{d}x \;=\; \left[s_{k+1}^{\mathbf{c}}(x)\right]_a^b \;=\; s_{k+1}^{\mathbf{c}}(\lambda_{g+1}) - s_{k+1}^{\mathbf{c}}(\lambda_0), \tag{37}$$

because $a=\lambda_0$, $b=\lambda_{g+1}$. Considering the definition, properties of B-splines and the mentioned additional knots (19) it follows that

$$s_{k+1}^{\mathbf{c}}(\lambda_{g+1}) - s_{k+1}^{\mathbf{c}}(\lambda_0) = c_g - c_{-k-1}. \tag{38}$$

Thus

$$\int_a^b s_k^{\mathbf{b}}(x)\,\mathrm{d}x \;=\; c_g - c_{-k-1}. \tag{39}$$

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 (36) 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 (35) we easily get

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

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

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

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

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

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

Proof of Theorem 6.1

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 B: 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 functional

$$J_l(s_k) = \alpha \int_a^b \left[s_k^{(l)}(x)\right]^2 \mathrm{d}x + \sum_{i=1}^{n} w_i \left[y_i - s_k(x_i)\right]^2 \tag{40}$$

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 formula

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

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

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

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

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

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

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

with

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

and

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

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 .

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., 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.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

18/2017 Ambartsumyan, I.; Khattatov, E.; Yotov, I.; Zunino, P.
A Lagrange multiplier method for a Stokes-Biot fluid-poroelastic structure interaction model

19/2017 Giovanardi, B.; Formaggia, L.; Scotti, A.; Zunino P.
Unfitted FEM for modelling the interaction of multiple fractures in a poroelastic medium

20/2017 Albrecht G.; Caliò F.; Miglio E.
Fair surface reconstruction through rational triangular cubic Bézier patches

16/2017 Ghiglietti, A.; Scarale, M.G.; Miceli, R.; Ieva, F.; Mariani, L.; Gavazzi, C.; Paganoni, A.M.; E
Urn models for response-adaptive randomized designs: a simulation study based on a non-adaptive randomized trial

15/2017 Tagliabue, A; Dede', L.; Quarteroni A.
Complex blood flow patterns in an idealized left ventricle: a numerical study

14/2017 Bruggi, M.; Parolini, N.; Regazzoni, F.; Verani, M.
Finite Element approximation of an evolutionary Topology Optimization problem

13/2017 Gigante, G.; Vergara, C.
Optimized Schwarz Methods for circular flat interfaces and geometric heterogeneous coupled problems

17/2017 Agosti, A.
Error Analysis of a finite element approximation of a degenerate Cahn-Hilliard equation

10/2017 Pini, A.; Stamm, A.; Vantini, S.
Hotelling's T^2 in separable Hilbert spaces

12/2017 Gasperoni, F.; Ieva, F.; Barbati, G.; Scagnetto, A.; Iorio, A.; Sinagra, G.; Di Lenarda, A.
Multi state modelling of heart failure care path: a population-based investigation from Italy