PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISERTATION THESIS

Bayes spaces and their applications

Supervisor: **doc. RNDr. Karel Hron, Ph.D.**
Author: **Mgr. Renáta Talská**
Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: Full-time
The year of submission: 2020

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Renáta Talská

Název práce: Bayesovy prostory a jejich aplikace

Typ práce: Disertační práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2020

Abstrakt: Hustotami rozdělení pravděpodobností (angl. probability density functions, PDFs) rozumíme funkcionální data nesoucí relativní informaci. Jejich vlastnosti jako invariantnost na změnu měřítka a relativní měřítko jsou zachyceny pomocí Bayesových prostorů měr; Bayesovy prostory tak představují zobecnění Aitchisonovy geometrie pro kompoziční data. Tyto prostory mají strukturu Hilbertova prostoru, jehož počátek je dán referenční mírou, která může být jednoduše změněna pomocí známého řetězového pravidla. Algebraická struktura Bayesových prostorů umožňuje PDFs vyjádřit jako reálné funkce ve standardním L^2 prostoru vzhledem ke zvolené referenční míře použitím centrované logpodílové (clr) transformace. Toto je klíčové pro možnost užití metod funkcionální analýzy dat (functional data analysis, FDA) pro statistické zpracování hustot, neboť tyto metody jsou typicky navržené právě v prostorech L^2. Protože výsledné transformované PDFs mají nulový integrál (vzhledem k dané referenční míře), jedná se o prvky podprostoru L^2, který je dále označen jako L^2_0. Cílem této disertační práce je představit Bayesovy prostory jako prostory hustot na omezeném intervalu s (i) Lebesgueovou a (ii) obecnou pravděpodobnostní referenční mírou, a jejich aplikace pro vybrané metody FDA. Podobně jako v FDA, vhodné statistické předzpracování diskrétně pozorovaných PDFs je klíčové pro jejich následnou analýzu. Nová metodika založená na principech Bayesových prostorů navrhuje užití (vyhlazovacích) splajnů nazvaných kompoziční (vyhlazující) splajny. Jejich konstrukce je založena na vytvoření B-splajnového bázového systému přímo v prostoru L^2_0 vzhledem k Lebesgueově referenční míře. Následně mohou být kompoziční splajny implementovány do FDA metod pro statistické zpracování PDFs, což je podrobně demonstrováno na případu regresní analýzy se závisle proměnnou reprezentovanou PDFs. Disertační práce se věnuje i aspektu vážení oboru hodnot hustot prostřednictvím referenční míry. Vliv změny referenční míry na statistickou analýzu PDFs je demonstrován pomocí funkcionální metody hlavních komponent na souboru dat o příjmech v Itálii. Pro její implementování, stejně tak jako dalších metod FDA, je klíčové použití nové clr transformace, která zobrazí Bayesovy prostory s obecnou referenční mírou do L^2_0 prostorů s Lebesgueovou referenční mírou.

Klíčová slova: Bayesovy prostory, hustoty rozdělení pravděpodobností, referenční míra, centrovaná logpodílová transformace, B-splajnová reprezentace, regresní analýza, funkcionální metoda hlavních komponent

Počet stran: 114

Počet příloh: 0

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Renáta Talská

Title: Bayes spaces and their applications

Type of thesis: Dissertation thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2020

Abstract: Probability density functions (PDFs) are understood as functional data carrying relative information. Their features such as scale invariance and relative scale are well captured by the theory of Bayes spaces of measures; Bayes spaces thus represent a generalization of the Aitchison geometry for compositional data. These spaces have a Hilbert space structure whose origin is determined by a given reference measure and it can be easily changed through the well-known chain rule. The algebraic-geometric structure of these spaces enables to express PDFs as real functions in the standard L^2 space with the same reference measure using the centered logratio (clr) transformation. This is key to propose statistical methods for PDFs by adapting popular methods of functional data analysis (FDA) which are typically designed in the L^2 space. Since the resulting transformed PFDs have the zero integral (with respect to the given reference measure), they are elements of a subspace of the L^2 space, hereafter denoted as L^2_0. The thesis aims to introduce Bayes spaces of PDFs on a bounded domain in case of (i) the Lebesgue measure and (ii) a general probability measure, and their application to selected problems of FDA. Similar as in FDA, a proper statistical preprocessing of discretely sampled PDFs is crucial for any further analysis. A novel methodology based on principles of Bayes spaces was developed and proposes to use (smoothing) spline functions called compositional (smoothing) splines. Their construction relies on building up a B-spline basis system directly in the L^2_0 space w.r.t. the Lebesgue reference measure. Consequently, the compositional splines can be implemented into FDA methods for statistical processing of PDFs, as demonstrated in detail in case of regression analysis with functional response formed by PDFs. The thesis further deals with weighting of PDFs through the reference measure. The impact on statistical analysis is illustrated through an application to the functional principal component analysis of Italian income data. For its implementation, as well as for the other methods of FDA, it is essential to use a novel centered logratio transformation that maps Bayes spaces with a general reference measure into the L^2_0 space with the Lebesgue reference measure.

Key words: Bayes spaces, probability density functions, reference measure, centered logratio transformation, compositional splines, B-spline representation, functional regression analysis, functional principal component analysis

Number of pages: 114

Number of appendices: 0

Language: English

Statement of originality

I hereby declare that this dissertation thesis has been completed independently, under the supervision of Doc. RNDr. Karel Hron, Ph.D. All the materials and resources are cited concerning scientific ethics, copyrights and the laws protecting intellectual property. This thesis or its parts were not submitted to obtain any other or the same academic title.

In Olomouc

Contents

Acknowledgment

I am particularly grateful to my supervisor Karel Hron for his continuous support of my Ph.D. study and related research, for his patience, motivation, and invaluable advice. I am also very thankful for the continuous assistance given by Alessandra Menafoglio and for her hospitality during my stay in Milan.

I would like to also express my very profound gratitude to my parents and to my partner for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. Finally, I want to thank my colleagues from Olomouc, especially to Julie, Nikola and Veronika, whose support and cooperation were immensely important to me as well.

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, a PDF is a non-negative Borel measurable function constrained to integrate to a constant, conventionally set to one. Several authors [9, 12, 42, 43] noted that PDFs have a *relative* nature, in the sense that the meaningful information is embedded in the relative contribution of the probability of any (Borel) subset of the domain of the random variable generating the data to the overall probability, i.e. the measure of the whole set (so-called *total*). Changing the value of the total by multiplying the PDF by a positive real constant results in a scaled density conveying the same *relative* information (which is known as the *scale invariance* property). As a consequence, the actual total is in fact irrelevant for the purpose of the analysis, as widely recognized in Bayesian statistics [19]. The total used simply determines a representative of the equivalence class of proportional density functions.

The relative nature of PDFs can be explained directly with an example: the relative increase of a probability over a Borel set from 0.05 to 0.1 (2 multiple) differs from the increase 0.5 to 0.55 (1.1 multiple), although the absolute differences are the same in both cases. This is known as the *relative scale* property of PDFs. It motivates the use of the so-called logratio approach – a well-established methodology for the analysis of compositional data. These are vectors describing quantitatively the parts of some whole, and are frequently represented as constrained data (e.g. proportions, percentages) carrying relative information [1, 33]. PDFs can be then interpreted as the continuous counterparts of compositions, i.e., as compositions with infinitely many parts. This has recently motivated the construction of the so-called Bayes Hilbert spaces, whose geometry results from the generalization of the Aitchison geometry for compositional data [14] to the infinite-dimensional case. While the pioneering work on Bayes spaces [12] covers only the case assuming that densities are defined on a finite support, Van den Boogaard et. al [43] extended this concept even for densities on possibly unbounded support and introduced Bayes spaces in a more general setting, i.e. as spaces of measures endowed with the Hilbert space structure. In general, Bayes spaces can be defined only if a reference measure P has been set. In the pioneering work, the reference measure was set by default to the Lebesgue (i.e. uniform) reference measure, however, to deal with unbounded support, a non-uniform reference measure P has to be considered as it shown in the latter work. Although Bayes spaces allow to deal with both unbounded and bounded domains for the PDFs, the latter case has been mainly considered so far in practice, and it will be the main focus in this thesis.

Nowadays, we experience an increasing interest in the development of statistical methods for the analysis of PDFs [4, 21, 22, 28, 30, 31, 34, 35, 37]. Although functional data analysis (FDA) [36] may potentially provide a wide range of methodological tools for this purpose, they are typically designed for data embedded in the L^2 space of square-integrable functions. As such, they can not be applied directly to densities since the metric of L^2 spaces does not account for their peculiar properties (e.g., the aforementioned scale invariance and relative scale). The key point in the analysis of PDFs is to map them from Bayes spaces to L^2 spaces where standard FDA methods (e.g., smoothing of functional data, clustering, regression analysis, functional principal component analysis, etc.) can be applied.

The thesis aims to introduce the concept of Bayes space methodology which turns out to be a relevant approach to statistical analysis of PDFs. Three particular novel approaches to statistical processing of PDFs will be presented such as smoothing of PDFs [24], functional regression with the response variable represented by PDF [41] and weighting in Bayes spaces with implications for dimensionality reduction of PDFs using simplicial functional principal component analysis [40].

The first part of the thesis (Section 1) introduces Bayes spaces and, as a special case, the Aitchison geometry for compositional data, both in the case of the Lebesgue reference measure. The concept of these spaces will be considered in a more general setting, that is as spaces of probability measures, discrete and continuous, on a finite and bounded domain, respectively.

In Section 2, so-called *centered logratio (clr) transformation* is recalled. It maps PDFs from Bayes spaces with the Lebesgue reference measure to L^2 spaces with the same reference measure, so that PDFs can be processed statistically in these L^2 spaces instead of the original spaces while their relative nature is still taken in account. Nevertheless, PDFs are represented through zero-integral elements of L^2 spaces. More precisely, they belong to L^2_0 space that consists of elements of the L^2 space with zero-integral. This constraint cannot be neglected in the statistical processing of clr transformed PDFs as well as in their preprocessing. Accordingly, to estimate the underlying continuous density from (discretized) distributional observations, the optimal smoothing splines using standard B-spline basis functions from L^2 spaces for clr transformed densities proposed by Machalová et al. [23] will be recalled. Moreover, a novel approach which deals with a new class of B-spline basis functions directly in L^2_0 is proposed; it leads to a definition of the so-called *compositional* B-*spline basis system* in Bayes spaces [24]. The section contains also a construction of smoothing compositional splines and possible orthonormalization of the compositional spline basis which might be useful in some applications.

The resulting spline representation is used for regression analysis in the presence of a distributional response, and it is discussed in Section 3. The key point of the proposed approach is to consider PDFs as elements of a Bayes space with the Lebesgue reference measure, and accordingly to deal with them by respecting the geometry of this space. The clr transformation is used to map the regression model from Bayes spaces to L^2 spaces which then ease the computations. A B-spline representation of clr transformed data 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 using bootstrap methods.

A statistical tool for weighting of a (bounded) domain of PDFs is proposed in Section 4. It is shown that a weighting scheme can be embedded into the Bayes spaces by setting up a non-uniform reference measure P. In this section, Bayes spaces are built upon the reference measure P on a bounded domain and their properties are discussed in detail. The clear guidelines for the use of non-uniform reference measures are given, and the consequences of changing the reference measure from uniform to non-uniform one are explored. A particular interest is devoted to the clr transformation for a general reference measure P, and to a novel *unweighting* clr transformation which maps Bayes spaces with the general reference measure P into L^2 spaces with the Lebesgue reference measure, so that it allows an adaptation of FDA methods to the Bayes space setting when the weighting of the domain of PDFs is considered.

In the final Section 5, the effect of the weighting on a statistical analysis of PDFs is demonstrated in the context of weighted simplicial functional principal component analysis (wSFPCA), which extends this statistical method designed originally for distributional data in case of the Lebesgue reference [21] to this more general setting. Its implementation is based on the mapping of the wSFPCA model into L^2 spaces with the Lebesgue reference measure via the proposed unweighting clr transformation.

This dissertation thesis is based on the following papers that were published, accepted or submitted during my Ph.D. study:

- M. Hošek, J. Pacina, J. Štojdl, O. Bábek, J. Sedláček, K. Hron, **R. Talská**, S. Křížinecká, J. Fikarová, T. Matys Grygar, Change in geochemistry of fluvial sediments after dam construction (the Chrudimka River, the Czech Republic). *Applied Geochemistry*, 98:94-108, 2018.

- J. Machalová, **R. Talská**, K. Hron, A. Gába, Compositional splines for representation of density functions (*under review*).

- **R. Talská**, A. Menafoglio, J. Machalová, K. Hron, E. Fišerová, Compositional regression with functional response. *Computational Statistics and Data Analysis*, 123:66-85, 2018.

- **R. Talská,**, A. Menafoglio, K. Hron, J. J. Egozcue, J. Palarea-Albaladejo, Weighting the domain of probability densities in functional data analysis (*Stat, accepted for publication*).

1 Bayes spaces

Bayes spaces represent an algebraic-geometric structure of equivalence classes of proportional σ-finite measures, including probability measures. An arbitrary σ-finite measure P can be selected as the origin of the space. Once such a measure is stated, all measures can be identified with density functions with respect to the measure P, resulting from considering densities as Radon-Nikodym derivatives. Accordingly, P is referred to as the reference measure. Although the framework of Bayes spaces of general measures (i.e. finite and infinite measures, with bounded or unbounded support) has been introduced by [43], its construction for the unbounded supports raises further issues – both methodological and practical – which are still open. For this reason, in the following we restrict our attention to measures on a bounded domain $\Omega = [a, b] \subset \mathbb{R}$, which was demonstrated to be of broad applicability by several authors [7, 21, 29, 30, 31, 41, 40]. For restricted domain Ω, both discrete and continuous probability measures can be considered. In this setting, the reference measure is set by default to a uniform measure, i.e. to the counting measure (discrete case) and the Lebesgue measure (continuous case). The choice of the reference measure other than the standard uniform one induces weighting effects on the domain of measures as we will see in Section 4. Accordingly, Bayes spaces with the uniform reference measures are referred to as *unweighted* Bayes spaces and those with the non-uniform reference measures to as *weighted* Bayes spaces. The section aims to summarize the basics of the Bayes space methodology and to familiarize the reader with its Hilbert space structure, mainly in the case of the Lebesgue reference measure.

1.1 Unweighted Bayes spaces: sample space

We assume that the distribution of a continuous random variable is characterized by a σ-finite positive measure μ on a measurable space $(\Omega, \mathcal{A})$ with a reference measure P, $\Omega = [a, b] \subset \mathbb{R}$ and $\mathcal{A}$ being Borel σ-algebra $B([a, b])$. In this setting, the reference measure P can be set to the Lebesgue measure λ, restricted here to a bounded support. The reference density is then the reference measure with respect to itself, i.e. $d\lambda/d\lambda = 1$ on Ω. Given two measures μ and ν with λ-densities $f = d\mu/d\lambda$ and $g = d\mu/d\lambda$, we say that two measures (densities) are $\mathcal{B}(\lambda)$-equivalent, denoted by $\nu =_{\mathcal{B}(\lambda)} \mu$ ($f =_{\mathcal{B}(\lambda)} g$), if they are proportional. That is, in terms of measures, if there exists a positive real constant c such that, for any subset $\mathrm{B} \in B([a, b])$, $\mu(\mathrm{B}) = c \cdot \nu(\mathrm{B})$. If $\mu(\Omega) = 1$ (i.e., μ is a probability measure), we single out a particular representative within a $\mathcal{B}(\lambda)$-equivalence class of proportional measures (densities) which provides the same *relative* information. Indeed, this is typically quantified through the (log-)ratios $\mu(\mathrm{B}_1)/\mu(\mathrm{B}_2)$, with B_1, B_2 in $B([a, b])$ (equivalently in terms of densities, i.e, $f(t_1)/f(t_2)$, with t_1, t_2 in $\Omega = [a, b]$), which are clearly invariant within the $\mathcal{B}(\lambda)$-equivalence class (i.e. *scale invariance* is followed). Within the concept of Bayes spaces, the only relevant information embedded into measures (densities) itself is the relative one. This motivated the use of the log-ratio approach, already known from (multivariate) compositional data analysis, to deal with density functions.

For a fixed reference measure $\mathsf{P} = \lambda$, Bayes space $\mathcal{B}^2(\lambda)$ is a space of $\mathcal{B}(\lambda)$-equivalence classes of σ-finite positive measures on $\Omega = [a, b]$ with square-integrable log-density with respect to reference measure λ:

$$\mathcal{B}^2(\lambda) = \left\{ \mu \in \mathcal{B}^2(\lambda) : \int \left| \ln \frac{d\mu}{d\lambda} \right|^2 d\,\lambda < +\infty \right\}, \tag{1}$$

where measures are identified with the corresponding Radon-Nikodym densities; or, equivalently, we can say that $\mathcal{B}^2(\lambda)$ consists of $\mathcal{B}(\lambda)$-equivalence classes of proportional density functions $f = \frac{d\mu}{d\lambda}$ on $\Omega = [a, b]$ whose logarithm is square-integrable w.r.t. λ. We note that $\mathcal{B}(\lambda)$ is a space for measures as well as for densities since in both cases they are elements of this space. Nevertheless, whether $\mathcal{B}(\lambda)$ is interpreted as space of densities or measures should be obvious from the context.

In case of discrete random variables, σ-finite positive measure μ is considered on measurable space $\Omega = \{t_1, \ldots, t_D\}$ which consists of D possible values (categories) of the variable, and the reference measure P is set to a counting measure P^c on Ω. The reference density is again the measure with respect to itself, $p^c = d\mathsf{P}^c/d\mathsf{P}^c = 1$ on Ω. The reference measure P^c assigns to each subset B of Ω number of elements of Ω contained in B. For instance, $\mathsf{P}^c(\{t_i\}) = 1$ and measure of whole Ω is D, i.e. $\mathsf{P}^c(\Omega) = \sum_{i=1}^{D} \mathsf{P}^c(\{t_i\}) = D$. The measure μ assigns *volume*

x_i to each t_i from Ω, which identifies its density (i.e. probability function). The density $f = d\mu/d\mathsf{P}_c$, i.e.,

$$f(t) = (f(t_1), \dots, f(t_D))' = (x_1, \dots, x_D)' = \mathbf{x},$$

can be viewed as D-part compositional vector (or composition for short) with strictly positive parts carrying relative information. That is, ratios between parts of a composition capture the relevant information and log-ratio approach should be applied for their statistical analysis. Bayes space $\mathcal{B}^2(\mathsf{P}^c)$ is now $D-1$ dimensional Euclidean space known as the Aitchison geometry. The sample space of compositional data consists of equivalence classes of proportional vectors (compositions) $\mathbf{x}$ with D parts summing up to a positive constant. If $\mu(\Omega) = 1$, the proportional representation of the respective equivalence class is obtained, being element of the unit simplex $\mathcal{S}^D$, defined as

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \dots, x_D)' \in \mathcal{S}^D : x_i > 0, i = 1, \dots, D; \sum_{i=1}^{D} x_i = 1 \right\}. \qquad (2)$$

1.2 Hilbert structure of unweighted Bayes spaces

In this section, we focus on the Hilbert space geometry of unweighted Bayes spaces $\mathcal{B}^2(\lambda)$. The basic operations named *perturbation* ($\oplus$) and *powering* ($\odot$) represent addition and multiplication in $\mathcal{B}^2(\lambda)$. Moreover, the first of them can be interpreted as Bayes updating which gave the name to these spaces. The operations are defined as follows,

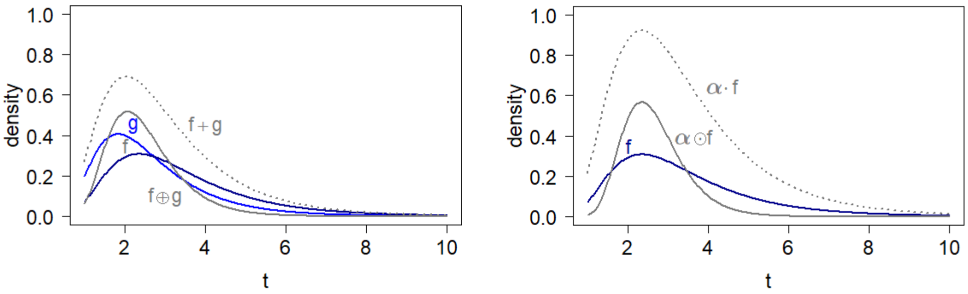
$$(\mu \oplus \nu)(\mathrm{B}) =_{\mathcal{B}(\lambda)} \int_{\mathrm{B}} \frac{d\mu}{d\lambda} \cdot \frac{d\nu}{d\lambda}\, d\lambda, \quad \mathrm{B} \in B, \qquad (3)$$

and

$$(\alpha \odot \mu)(\mathrm{B}) =_{\mathcal{B}(\lambda)} \int_{\mathrm{B}} \left(\frac{d\mu}{d\lambda}\right)^{\alpha} d\lambda, \quad \mathrm{B} \in B; \qquad (4)$$

where μ and ν are measures in $\mathcal{B}^2(\lambda)$ and α is a real number. The operations (3) and (4) can be equivalently expressed using densities. That is, for $f = \frac{d\mu}{d\lambda}$ and $g = \frac{d\nu}{d\lambda}$ we have that

$$(f \oplus g)(t) =_{\mathcal{B}(\lambda)} f(t) \cdot g(t) \quad \text{and} \quad (\alpha \odot f)(t) =_{\mathcal{B}(\lambda)} f(t)^{\alpha}, \quad t \in \Omega. \qquad (5)$$



(a) Perturbation

(b) Powering

Figure 1: Comparison of basic operations in $\mathcal{B}^2(\lambda)$ and $L^2(\lambda)$. In panel (a), $f \oplus g$ indicates perturbation (grey line) and $f + g$ addition of two densities f and g (grey dotted line). In panel (b), $\alpha \odot f$ indicates powering (grey line) and $\alpha \cdot f$ multiplication of density f by real constant α (grey dotted line) with $\alpha = 3$.

In [12], it is proven that $\mathcal{B}^2(\lambda)$ equipped with the operations $(\oplus, \odot)$ is a vector space. Note that the neutral elements of perturbation and powering are $e(t) = \frac{1}{\lambda(\Omega)} = \frac{1}{b-a}$ (i.e., the uniform density), and 1, respectively. The operation subtraction $(\ominus)$ between two densities f, g is obtained as perturbation of f with reciprocal of g,

$$(f \ominus g)(t) =_{\mathcal{B}(\lambda)} f(t) \oplus [(-1) \odot g(t)] \quad t \in \Omega. \tag{6}$$

This operation is identified as the Radon–Nikodym derivative of μ with respect to ν, that is $\frac{d\mu}{d\lambda} \cdot \left(\frac{d\nu}{d\lambda}\right)^{-1} = \frac{d\mu}{d\nu}$. The results of operations (3), (4) and (6) are densities again, possibly rescaled to unit integral constraint using the closure operation $\mathcal{C}(f) = \frac{f}{\int_\Omega f d\lambda}$. Nevertheless, any other representatives of equivalence classes could be considered as well.

Figure 1 illustrates the effects of perturbation and powering compared to the standard operations of sum and product by a constant in $L^2(\lambda)$. We recall that the representatives of the equivalence classes of $\mathcal{B}^2(\lambda)$ are densities which integrate to unity. The perturbation of a density f by a density g $(f \oplus g)$ and the addition of f and g $(f + g)$ is represented in Figure 1a. We see that summation operation according to the geometrical structure of $L^2(\lambda)$ is not a probability density function, and therefore it is not appropriate as an operation in $\mathcal{B}^2(\lambda)$, whereas the perturbation operation results in a probability density function. The argument that supports this is that f and g are in fact (truncated) log-normal densities and as such, they belong to the exponential family which is closed with respect to the perturbation operation. More precisely, an exponential family is a finite dimensional affine subspace of the Bayes space $\mathcal{B}^2(\lambda)$ on Ω [42]. The density f is updated by the information contained in the density g (i.e. Bayesian updating of the information), so that the resulting density $f \oplus g$ is more concentrated and shifted towards g.

The second operation, powering of the density f by α ($\alpha \odot f$) and multiplication of f by α ($\alpha \cdot f$) with $\alpha = 3$ is displayed in Figure 1b. Firstly, notice that the multiplication in sense of the L^2 geometry leads to a scaled f – function, which is not probability density function – and as an operation in $\mathcal{B}^2(\lambda)$, it only changes the representative within the equivalence class. Accordingly, $\alpha \cdot f$ as the multiplication operation in $\mathcal{B}^2(\lambda)$ does not make any sense in this setting. Instead, $\alpha \odot f$ results in a density which is more concentrated, thus having lower variability: f is updated by f itself and subsequently density $f \cdot f$ is updated by f again. That is, it can be similarly linked to Bayesian updating of the information as in the previous case.

Finally, to endow $\mathcal{B}^2(\lambda)$ with the Hilbert space structure, an inner product is required. Egozcue et al. [12] defined it for the Lebesgue reference measure on $\Omega = [a, b]$ and van den Boogaard et al. [43] extended the definition to any finite reference measure. Accordingly, the Bayes inner product on $\mathcal{B}^2(\lambda)$ can be defined [40] as

$$\langle f, g \rangle_{\mathcal{B}(\lambda)} = \frac{1}{2\lambda(\Omega)} \int_\Omega \int_\Omega \ln \frac{f(t)}{f(u)} \ln \frac{g(t)}{g(u)} \, d\lambda(t) d\lambda(u), \quad t, u \in \Omega, \tag{7}$$

where $\lambda(\Omega) = b - a$, and the corresponding norm and distance as

$$\|f\|_{\mathcal{B}(\lambda)} = \sqrt{\langle f, f \rangle_{\mathcal{B}(\lambda)}} \quad \text{and} \quad d_{\mathcal{B}(\lambda)}(f, g) = \|f \ominus g\|_{\mathcal{B}(\lambda)} . \tag{8}$$

In the discrete setting, P is set to counting measure P^c and the basic operations (3) and (4) read

$$f \oplus g =_{\mathcal{B}(\mathsf{P}^c)} (f(t_1)g(t_1), \ldots, f(t_D)g(t_D))' \tag{9}$$

and

$$\alpha \odot f =_{\mathcal{B}(\mathsf{P}^c)} (f(t_1)^\alpha, \ldots, f(t_D)^\alpha)'\,; \tag{10}$$

where f and g are compositions from $\mathcal{S}^D$ and α is a real number. The output of the operations (9) and (10) is a composition, possibly closed to unit sum via closure operation, acting on the composition f as $\mathcal{C}(f) = \frac{f}{\sum_{i=1}^{D} f(t_i)}$. The inner product (7) reduces to

$$\langle f, g \rangle_{\mathcal{B}(\mathsf{P}^c)} = \frac{1}{2\mathsf{P}^c(\Omega)} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{f(t_i)}{f(t_j)} \cdot \ln \frac{g(u_i)}{g(u_j)}, \tag{11}$$

where $\mathsf{P}^c(\Omega) = D$, and the corresponding norm and distance to

$$\|\mathbf{f}\|_{\mathcal{B}(\mathsf{P}^c)} = \sqrt{\langle f, f \rangle_{\mathcal{B}(\mathsf{P}^c)}} \quad \text{and} \quad d_{\mathcal{B}(\mathsf{P}^c)}(f, g) = \|f \ominus g\|_{\mathcal{B}(\mathsf{P}^c)}\,. \tag{12}$$

Although van den Boogaart et al. [43] has shown that the geometry on the simplex $\mathcal{S}^D$ is a particular case of Bayes spaces with the reference measure P^c, the algebraic-geometric structure on $\mathcal{S}^D$ was initially proposed simultaneously already in [3] and [32], and it is commonly known as a *Aitchison geometry*.

2 First steps for a statistical analysis in unweighted Bayes spaces

As a first step of any data analysis, one needs to think about the sample space for data embedding. Whereas the natural choice in the case of multivariate observations is the Euclidean real space, Hilbert spaces are mostly employed for functional data due to their geometric structure which allows an easier extension of multivariate methods. In fact, the Hilbert space generalizes the concept of the Euclidean space to spaces of any (even infinite) dimension. For instance, statistical methods provided by the FDA are mostly developed under the assumption that the data belongs to the Hilbert space $L^2(\lambda)$ of squared-integrable functions with the reference measure defaulty set to the Lebesgue measure λ. Although discrete density functions can be viewed as multivariate observations and continuous density functions as functional data, they both share the property of scale invariance and relative scale which are not honored either by standard multivariate methods nor by the FDA. Nevertheless, as long as the data are embedded in a separable Hilbert space [12], an isometric mapping can be found which enables to express elements of Bayes spaces as real vectors of the Euclidean space or real functions of the $L^2(\lambda)$ space, respectively. In fact, all Hilbert spaces are isometric to each other [44]. Subsequently, standard statistical analysis can be performed via multivariate analysis or FDA, respectively, while accounting for the Bayes space geometry. Such mapping can be provided by centered logratio (clr) transformation and will be introduced and demonstrated in this section, mainly for the continuous case. Moreover, this transformation is the key to propose smoothing splines designed for clr transformed density functions which will also be introduced in this section.

2.1 Centered logratio transformation

Let us first focus on the continuous case, i.e. $\mathsf{P} = \lambda$. The clr mapping represents an isometric isomorphism (i.e. a bijective map preserving distances) between $\mathcal{B}^2(\lambda)$ and $L^2(\lambda)$ spaces and it is defined in [43] for $f \in \mathcal{B}^2(\lambda)$ as

$$f^c(t) = \mathrm{clr}_\lambda(f)(t) = \ln f(t) - \frac{1}{\lambda(\Omega)} \int_\Omega \ln f(u)\, d\lambda(u), \quad t \in \Omega. \tag{13}$$

Apparently, the clr representation allows to use the ordinary geometry of $L^2(\lambda)$ to conduct operations of perturbation (3), powering (4) and inner product (7) for the elements of $\mathcal{B}^2(\lambda)$, while accounting for the specific features captured by the Bayes space. Indeed,

$$\mathrm{clr}_\lambda(f \oplus g) = \mathrm{clr}_\lambda(f)(t) + \mathrm{clr}_\lambda(g)(t), \quad \mathrm{clr}_\lambda(\alpha \odot f)(t) = \alpha \cdot \mathrm{clr}_\lambda(f)(t) \tag{14}$$

and

$$\langle f, g \rangle_{\mathcal{B}^2(\lambda)} = \langle \mathrm{clr}_\lambda(f), \mathrm{clr}_\lambda(g) \rangle_{L^2(\lambda)} = \int_\Omega \mathrm{clr}_\lambda(f)(t) \cdot \mathrm{clr}_\lambda(g)(t) d\lambda(t), \tag{15}$$

where $\langle \cdot, \cdot \rangle_{L^2(\lambda)}$ denotes the inner product in $L^2(\lambda)$. However, due to the construction, the clr transformed densities are characterized by zero-integral constraint (w.r.t. λ),

$$\int_\Omega \mathrm{clr}_\lambda(f)(t)\, d\lambda(t) = \int_\Omega \ln f(t)\, d\lambda(t) - \int_\Omega \frac{1}{\lambda(\Omega)} \int_\Omega \ln f(u)\, d\lambda(u)\, d\lambda(t) = 0, \tag{16}$$

which needs to be taken into account when analyzing clr transformed densities. As the clr space is clearly a subspace of $L^2(\lambda)$, hereafter it is denoted as $L^2_0(\lambda)$. Note that clr transformation represents one-to-one mapping, so it is possible to map densities in $L^2_0(\lambda)$ back to $\mathcal{B}^2(\lambda)$ by using exponential transformation, i.e. $\exp\left[f^c\right](t) = \exp\left[\mathrm{clr}_\lambda(f)\right](t)$. The resulting back-transformed density f can be closed to the unit integral due to the scale invariance feature.

In the discrete case ($\mathsf{P} = \mathsf{P}^c$), the clr transformation (13) reads

$$\begin{aligned} f^c(t) = \mathrm{clr}_{\mathsf{P}^c}(f)(t) &= \ln f(t) - \frac{1}{\mathsf{P}^c(\Omega)} \sum_{i=1}^{D} \ln f(u_i) = \ln f - \ln(f(u_1) \cdot \ldots \cdot f(u_D))^{\frac{1}{D}} \\ &= \left(\ln \frac{f(t_1)}{\mathrm{gm}(f)}, \ldots, \ln \frac{f(t_D)}{\mathrm{gm}(f)} \right)' \end{aligned} \tag{17}$$

where $\mathrm{gm}(f) = \left(\prod_{i=1}^{D} f(u_i) \right)^{\frac{1}{D}}$ stands for the geometric mean of D components and $t, u \in \Omega$. The discrete version of the clr transformation was initially proposed by [1] and from the geometric point of view it corresponds to coefficients with respect to a generating system on simplex. Similarly, it translate basic operations and metrics of the $\mathcal{B}(\mathsf{P}^c)$ into the usual operations and metrics of a standard Euclidean space. That is,

$$\mathrm{clr}_{\mathsf{P}^c}(f \oplus g) = \mathrm{clr}_{\mathsf{P}^c}(f) + \mathrm{clr}_{\mathsf{P}^c}(g), \quad \mathrm{clr}_{\mathsf{P}^c}(\alpha \odot f) = \alpha \cdot \mathrm{clr}_{\mathsf{P}^c}(f)$$

and

$$\langle f, g \rangle_{\mathcal{B}^2(\mathsf{P}^c)} = \langle \mathrm{clr}_{\mathsf{P}^c}(f), \mathrm{clr}_{\mathsf{P}^c}(g) \rangle_2 \,, \tag{18}$$

with $\langle \cdot, \cdot \rangle_2$ denoting inner product of a standard Euclidean space. Note that sum of clr coefficients is zero, making the transformation (17) of limited use in many statistical methods. The way out is to express a composition f via orthonormal coordinates called balances [14], i.e. as Fourier coefficients of a basis in $\mathcal{B}(\mathsf{P}^c)$. The inverse mapping to the clr transformation to $\mathcal{B}(\mathsf{P}^c)$ is obtained by using the exponential, $\exp\left[f^c\right] = \exp\left[\mathrm{clr}_{\mathsf{P}^c}(f)\right]$, possibly closed to the unit sum of parts of the resulting composition.

2.1.1 Effect of clr transformation

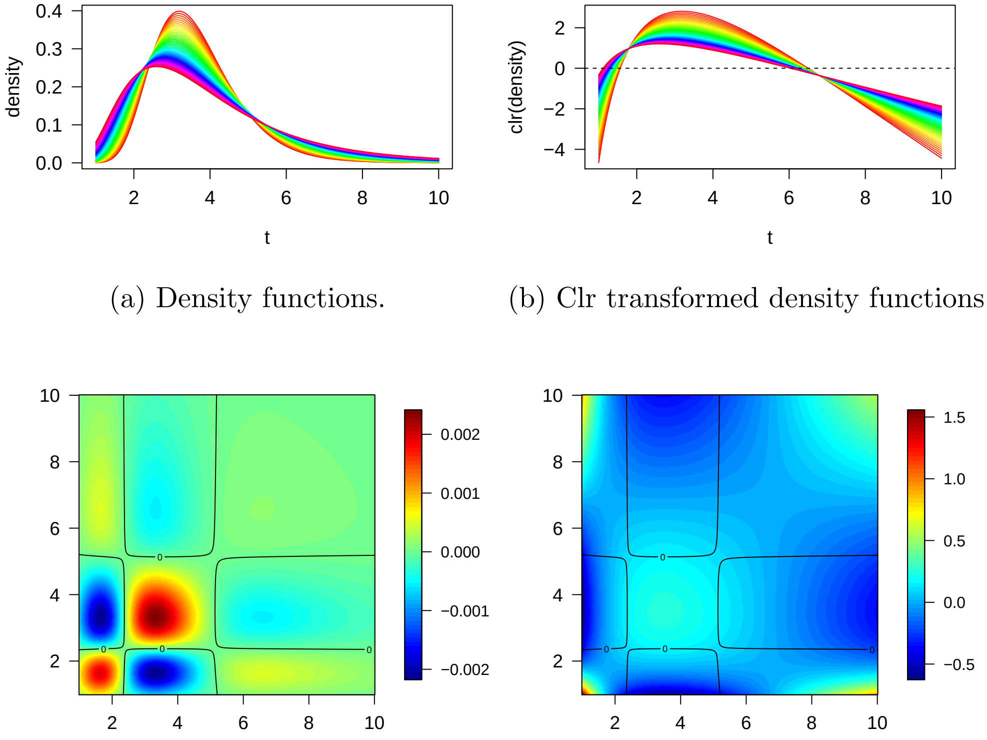
In order to examine the effects of the clr transformation (13), we simulate densities from the exponential family, namely log-normal family of distributions. We consider a set of (truncated) log-normal densities with the mean $\mu = 1.25$ and standard deviations $\sigma_j = 0.3 + 0.005 \cdot (j-1)$ for $j = 1, \ldots, 50$, on the interval $\Omega = [1, 10]$. Their representation with respect to the Lebesgue measure,

$$f_\lambda(t; \sigma_j) =_{\mathcal{B}(\lambda)} \frac{1}{t \cdot \sigma_j} \exp\left\{ -\frac{(\ln t - 1.25)^2}{2\sigma_j^2} \right\}, \; t \in \Omega, \tag{19}$$

is displayed on colored scale (Figure 2a) together with their clr transforms (Figure 2b), obtained as

$$f_\lambda^c(t; \sigma_j) = -\frac{\ln^2 t}{2\sigma_j^2} + \left(-1 + \frac{1.25}{\sigma_j^2}\right)\left(\ln t - \frac{10}{9} \cdot \ln 10 + 9\right) + \frac{1}{\sigma_j^2}\left(1 + \frac{5}{9} \cdot \ln^2 10 - \frac{9}{10} \ln 10\right), t \in \Omega. \tag{20}$$

Since densities are functions characterized by their relative nature and this feature is followed by Bayes spaces, the main variability in data set is present on



(a) Density functions.

(b) Clr transformed density functions.

(c) Covariance function in $L^2(\lambda)$.

(d) Covariance function in $\mathcal{B}^2(\lambda)$.

Figure 2: Log-normal density functions w.r.t. the Lebesgue measure with parameters $\mu = 1.25$ and $\sigma_j = 0.3 + 0.005 \cdot (j-1)$ for $j = 1, \ldots, 9$, $\Omega = [1, 10]$.

boundaries of the domain Ω, being more dominant in its left-hand side (see Figure 2a). It corresponds to the fact, that densities have larger relative differences in their values here. This is well reflected by clr transformation in Figure 2b where the variability is driven by absolute differences among curves. Note that densities are mapped into $L_0^2(\lambda)$ via clr transformation, so that they can be interpreted in agreement with the L^2 space geometry considered therein. On the other hand, if $L^2(\lambda)$ is chosen for processing of the original densities (Figure 2a), their relative nature is completely ignored since the variability is exhibited for parts of the domain with higher absolute values of the curves. This is apparent when comparing the covariance functions estimated in $L^2(\lambda)$ and in $\mathcal{B}^2(\lambda)$ spaces (Figure 2c and 2d, respectively).

2.2 Smoothing of density functions

As demonstrated in the previous section, as the first step of any data analysis, the sample space for the embedding data needs to be chosen correctly to avoid misleading results which would not reflect the actual properties of data. Since functional data rarely occur in practice in their continuous form, the second step in FDA is related to the estimation of the underlying N functions $f_1, \dots, f_N$ from discretized data $(t_{ij}, y_{ij}), i = 1, \dots, N, j = 1, \dots, n_i$, where y_{ij} is observation of f_i at t_{ij}. In case of density functions, they are discretely sampled in terms of histogram data. That is, for each density $f_i(t), t \in \Omega, i = 1, \dots, N$, one usually observes a positive real vector $\mathbf{W}_i = (W_{i1}, \dots, W_{iD_i})'$, whose components correspond to the (absolute or relative) frequencies of D_i classes in which the interval Ω is partitioned. Accordingly, the raw density data y_{ij} correspond to interval midpoints t_{ij} of D_i classes obtained by dividing (not necessary normalized) components of $\mathbf{W}_i$ by the length of the respective classes. Note that data $\mathbf{y}_i = (y_{i1}, \dots, y_{iD_i})', i = 1, \dots, N$ can be interpreted as discretized density functions, that is, as compositions. Since it is convenient to perform preprocessing of density functions in clr space L_0^2, discrete clr transformation (17) is employed to express compositional vectors $\mathbf{y}_i, i = 1, \dots, N$ in a standard Euclidean space; this yields clr transformed data denoted as $\mathbf{z}_i = (z_{i1}, \dots, z_{iD_i})', i = 1, \dots, N$. Consequently, the aim is to estimate (approximate) the underlying continuous clr density functions $\mathrm{clr}_\lambda(f_i), i = 1, \dots, N$ from raw given data $(t_{ij}, z_{ij}), i = 1, \dots, N, j = 1, \dots, D_i$.

Spline functions are extensively used in FDA for an approximation of nonperiodical functions as they are flexible enough to cover a wide range of their specific behavior, hence they are also a natural choice for density functions. A first attempt of constructing a spline representation adapted for clr transformed density functions was proposed in [23]. The problem is that B-splines that form the basis system for the spline expansion come from L^2, but not from L_0^2. Therefore, an important step ahead is made by constructing a B-spline basis directly in the clr space L_0^2. As a direct consequence, the B-splines can be expressed directly in Bayes spaces leading to spline representation of density functions in the original space; hereafter we refer to *compositional splines.* Apart from methodological advantages, using compositional splines simplifies the construction and interpretation of spline coefficients that can be considered as coefficients of a basis in Bayes spaces. Moreover, using splines for the representation of density functions turned out to be the most appropriate approximative tool as the associated basis coefficients can be directly used for further statistical analysis, i.e., for instance in functional regression as we will see in Section 3. Additionally, we note that unlike to former approach (Approach I) which aims to develop splines honoring the zero integral constraint using the standard B-spline basis, the later one (Approach II) enables to implement the methods of FDA directly in Bayes spaces.

Both methods are detailed in the following, and it is assumed that a single density function $\mathrm{clr}_\lambda(f)$ is being approximated.

2.2.1 Approach I

First, we recall some basics related to spline theory. To set the notation, call values

$$\Delta\lambda := \{\lambda_0 = a < \lambda_1 < \ldots < \lambda_g < b = \lambda_{g+1}\} \tag{21}$$

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 $\Omega = [a,b]$ given the knots $\Delta\lambda$. It is known that $\dim(\mathcal{S}_k^{\Delta\lambda}[a,b]) = g+k+1$. For the construction of all basis functions of $\mathcal{S}_k^{\Delta\lambda}[a,b]$, it is necessary to consider some additional knots. Without loss of generality, we here assume that those additional knots are at the boundary, i.e.,

$$\lambda_{-k} = \cdots = \lambda_{-1} = \lambda_0, \quad \lambda_{g+1} = \lambda_{g+2} = \cdots = \lambda_{g+k+1}. \tag{22}$$

Then every spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ in the L^2 space has a unique representation as (see [6], [8] for details)

$$s_k\left(t\right) = \sum_{i=-k}^{g} b_i B_i^{k+1}\left(t\right), \tag{23}$$

where the vector $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ is the vector of B-spline basis coefficients of $s_k\left(t\right)$ and functions $B_i^{k+1}(t)$, $i = -k, \ldots, g$ are B-spline functions of the same degree k as spline function $s_k\left(t\right)$ forming basis in $\mathcal{S}_k^{\Delta\lambda}[a,b]$. They are defined for

$k = 0$ (order 1) by

$$B_i^1(t) = \begin{cases} 1 & \text{if } t \in [\lambda_i, \lambda_{i+1}) \\ 0 & \text{otherwise.} \end{cases}$$

and for k, $k \in \mathbb{N}$, $k \geq 1$, (order $k+1$) by

$$B_i^{k+1}(t) = \frac{t - \lambda_i}{\lambda_{i+k} - \lambda_i} B_i^k(t) + \frac{\lambda_{i+k+1} - t}{\lambda_{i+k+1} - \lambda_{i+1}} B_{i+1}^k(t).$$

In [23], the optimal smoothing problem was represented as a trade-off between smoothing and the least squares approximation. Assume that data (t_j, z_j), $a \leq t_j \leq b$, the weights $w_j^s \geq 0$, $j = 1, \dots, D$, $D \geq g+1$ and the parameter $\alpha \in (0,1)$ are given. For an arbitrary $l \in \{1, \dots, k-1\}$ our aim is to find a smoothing spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$, which minimizes the functional

$$J_l(s_k) = \alpha \sum_{j=1}^{D} w_j \left[z_j - s_k(t_j)\right]^2 + (1-\alpha) \int_a^b \left[s_k^{(l)}(t)\right]^2 dt, \tag{24}$$

and fulfills the condition

$$\int\limits_a^b s_k(t)\, dt = 0, \tag{25}$$

resulting from the clr transformation. The minimization problem (24) represents a compromise between staying close to the given data and obtaining a smooth function. The smoothness of the resulting approximation is affected by the smoothing parameters α and l, where l stands for lth derivative. Similarly, one can minimize the following functional with respect to condition (25) for some positive parameter α, i.e.

$$J_l(s_k) = \sum_{j=1}^{D} w_j \left[z_j - s_k(t_j)\right]^2 + \alpha \int_a^b \left[s_k^{(l)}(t)\right]^2 dt \tag{26}$$

For the sake of brevity, we will focus on the minimization of the functional (26). It was proven that the *optimal* smoothing spline for this task is the spline $s_k^*(t)$, given by formula

$$s_k^*\left(t\right) = \sum_{i=-k}^{g} b_i^* B_i^{k+1}\left(t\right), \tag{27}$$

with B-spline coefficients $\mathbf{b}^* = (b^*_{-k}, \ldots, b^*_g)'$ obtained as (see [23] for details)

$$\mathbf{b}^* = \mathbf{V}\mathbf{z} \tag{28}$$

with

$$\mathbf{V} := \mathbf{DK}\left[\alpha\left(\mathbf{DK}\right)'\mathbf{N}_{kl}\mathbf{DK} + \left(\mathbf{B}_{k+1}(\mathbf{t})\mathbf{DK}\right)'\mathbf{W}^s\mathbf{B}_{k+1}(\mathbf{t})\mathbf{DK}\right]^+\mathbf{K}'\mathbf{D}\mathbf{B}'_{k+1}(\mathbf{x})\mathbf{W}^s. \tag{29}$$

Here $\mathbf{A}^+$ denotes the Moore-Penrose pseudoinverse of a matrix $\mathbf{A}$, $\mathbf{W}^s = \operatorname{diag}(\mathbf{w}^s)$, $\mathbf{w}^s = (w^s_1, \ldots, w^s_D)'$, $\mathbf{t} = (t_1, \ldots, t_D)'$, $\mathbf{z} = (z_1, \ldots, z_D)'$,

$$\mathbf{B}_{k+1}(\mathbf{t}) = \begin{pmatrix} B^{k+1}_{-k}(t_1) & \ldots & B^{k+1}_g(t_1) \\ \vdots & \ddots & \vdots \\ B^{k+1}_{-k}(t_D) & \ldots & B^{k+1}_g(t_D) \end{pmatrix} \in \mathbb{R}^{D,g+k+1} \tag{30}$$

is the collocation matrix,

$$\mathbf{D} = (k+1)\operatorname{diag}\left(\frac{1}{\lambda_1 - \lambda_{-k}}, \ldots, \frac{1}{\lambda_{g+k+1} - \lambda_g}\right) \in \mathbb{R}^{g+k+1,g+k+1} \tag{31}$$

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

$$\mathbf{M}_{kl} = \begin{pmatrix} \left\langle B^{k+1-l}_{-k+l}, B^{k+1-l}_{-k+l}\right\rangle_{L^2(\lambda)} & \cdots & \left\langle B^{k+1-l}_{g}, B^{k+1-l}_{-k+l}\right\rangle_{L^2(\lambda)} \\ \vdots & & \vdots \\ \left\langle B^{k+1-l}_{-k+l}, B^{k+1-l}_{g}\right\rangle_{L^2(\lambda)} & \cdots & \left\langle B^{k+1-l}_{g}, B^{k+1-l}_{g}\right\rangle_{L^2(\lambda)} \end{pmatrix} \in \mathbb{R}^{g+k+1-l,g+k+1-l}, \tag{32}$$

where

$$\left\langle B^{k+1-l}_i, B^{k+1-l}_j\right\rangle_{L^2(\lambda)} = \int_a^b B^{k+1-l}_i(t)B^{k+1-l}_j(t)\,dt$$

stands for scalar product of B-splines in $L^2(\lambda)$ space. The matrix $\mathbf{S}_l$ is an upper triangular matrix such that

$$\mathbf{S}_l = \mathbf{D}_l\mathbf{L}_l \ldots \mathbf{D}_1\mathbf{L}_1 \in \mathbb{R}^{g+k+1-l,g+k+1}, \tag{33}$$

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

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

As an element of innovation, we aim to find the necessary and sufficient condition for the vector $\mathbf{b} = (b_{-k},\dots,b_g)'$ to be the vector of B-spline coefficients for spline with zero integral. The following Theorem 2.1 characterizes all the splines with zero integral (not necessarily a smoothing spline) using a standard B-spline basis system through a necessary and sufficient condition on $\mathbf{b}$.

Theorem 2.1 *For a spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$, $s_k(t) = \sum\limits_{i=-k}^{g} b_i B_i^{k+1}(t)$, the condition $\int\limits_a^b s_k(t)\,dt = 0$ is fulfilled if and only if $\sum\limits_{i=-k}^{g} b_i\,(\lambda_{i+k+1} - \lambda_i) \;=\; 0.$*

The proof of Theorem 2.1 is provided in Appendix A. It is easy to see that vector $\mathbf{b}$ is orthogonal to the vector $(\lambda_1 - \lambda_{-k},\dots,\lambda_{g+k+1} - \lambda_g)'$, that only depends on the knot positions. The following algorithm describes the computation of coefficients of any B-spline representation that fulfills the zero integral constraint.

Algorithm for finding a spline with zero integral

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

1. Choose $g+k$ arbitrary B-spline coefficients $b_i \in \mathbb{R}$, $i = -k\dots,j-1,j+1,\dots,g$,

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

$$\sum_{i=-k}^{g} b_i \left(\lambda_{i+k+1} - \lambda_i\right) \; = \; 0 \tag{34}$$

is fulfilled, and, with respect to Theorem 2.1, the spline $s_k\left(t\right) = \sum\limits_{i=-k}^{g} b_i B_i^{k+1}\left(t\right)$ satisfies condition $\int\limits_a^b s_k(t)\, dt = 0$.

Example 2.1 *We consider cubic spline, that is $k = 3$, and knots $\lambda_0 = 0 = a < 2 < 5 < 9 < 14 < 20 = b = \lambda_5$. It is obvious that $g = 4$. Additional knots are*

$$\lambda_{-3} = \lambda_{-2} = \lambda_{-1} = \lambda_0 = 0, \qquad 20 = \lambda_5 = \lambda_6 = \lambda_7 = \lambda_8.$$

We set $g + k = 7$ B-spline coefficients, for instance

$$b_{-3} = -4,\; b_{-2} = -4,\; b_{-1} = -3,\; b_0 = -3,\; b_1 = -2,\; b_2 = -1,\; b_4 = -2$$

Then we compute

$$b_3 = \frac{-1}{\lambda_7 - \lambda_3} \sum_{\substack{i=-3 \\ i \neq 3}}^{4} b_i \left(\lambda_{i+4} - \lambda_i\right) = 10.4$$

The spline with zero integral on interval $[0, 20]$ is given by formula

$$s_3\left(t\right) = \sum_{i=-3}^{4} b_i B_i^4\left(t\right)$$

where $\boldsymbol{b} = (-4, -4, -3, -3, -2, 10.4, -1, -2)'$ and $B_i^4(t), i = -3, \ldots, 4$ are B-spline basis functions, see Figure 3.

The resulting smoothing spline can be back-transformed into $\mathcal{B}^2(\lambda)$ using an exponential to see the approximation of the underlying density function in the original space. Nevertheless, we already mentioned that this methodology uses the standard B-spline system that belongs to $L^2(\lambda)$, not to $L_0^2(\lambda)$. The next subsection introduces so-called compositional spline functions which honor the zero integral constraint in $L_0^2(\lambda)$ for both the B-spline basis functions and the resulting spline function.

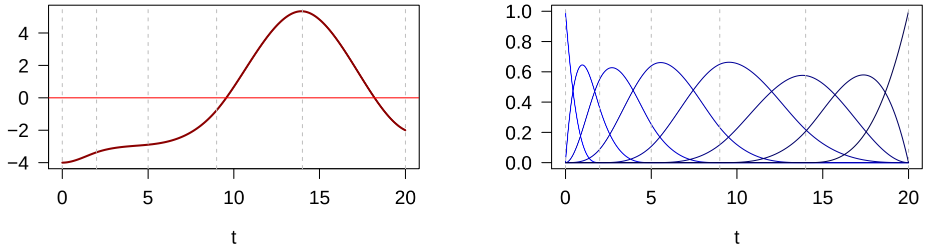


Figure 3: Cubic spline function $s_3(t)$ with zero integral in $\mathcal{S}_3^{\Delta\lambda}[0,20]$ (left) and B-spline basis system of $\mathcal{S}_3^{\Delta\lambda}[0,20]$ (right). Vertical dashed gray lines indicate knot positions.

2.2.2 Approach II

Let the sequence of knots (21) is given. We define the functions $Z_i^{k+1}(t)$ for $k \geq 0$, $k \in \mathbb{N}$, which are the first derivatives of the B-splines $B_i^{k+2}(t)$ for $k \geq 0$, $k \in \mathbb{N}$, as

$$Z_i^{k+1}(t) := \frac{d}{dt} B_i^{k+2}(t), \tag{35}$$

i.e., more precisely for $k = 0$

$$Z_i^1(x) = \begin{cases} 1 & \text{if } x \in [\lambda_i, \lambda_{i+1}) \\ -1 & \text{if } x \in (\lambda_{i+1}, \lambda_{i+2}] \end{cases}$$

and for $k \geq 1$

$$Z_i^{k+1}(t) = (k+1)\left(\frac{B_i^{k+1}(t)}{\lambda_{i+k+1} - \lambda_i} - \frac{B_{i+1}^{k+1}(t)}{\lambda_{i+k+2} - \lambda_{i+1}}\right). \tag{36}$$

The functions $Z_i^{k+1}(t)$ have similar properties as B-spline functions $B_i^{k+1}(t)$. They are piecewise polynomials of degree k on local support for $k \geq 1$,

$$\text{supp } Z_i^{k+1}(t) \;=\; \text{supp } B_i^{k+2}(t) = [\lambda_i, \lambda_{i+k+2}],$$

with continuous derivatives up to order $k-1$. From the perspective of L_0^2 space, a crucial point is that the integral of $Z_i^{k+1}(t)$ equals to zero. If we consider Curry-Schoenberg B-spline $M_i^{k+1}(t)$ [6], which are defined as

$$M_i^{k+1}(t) \;:=\; \frac{k+1}{\lambda_{i+k+1} - \lambda_i}\, B_i^{k+1}(t)$$

with property

$$\int\limits_{\mathbb{R}} M_i^{k+1}(t)\,dt \;=\; 1,$$

than it is clear that

$$Z_i^{k+1}(t) \;=\; M_i^{k+1}(x) - M_{i+1}^{k+1}(t) \tag{37}$$

and

$$\int\limits_{\mathbb{R}} Z_i^{k+1}(t)\,dt \;=\; 0.$$

In Figure 4, an instance of piecewise linear, quadratic and cubic polynomials $Z_i^{k+1}(t)$ is plotted; it visually confirms that zero-integral condition of these functions is fulfilled.

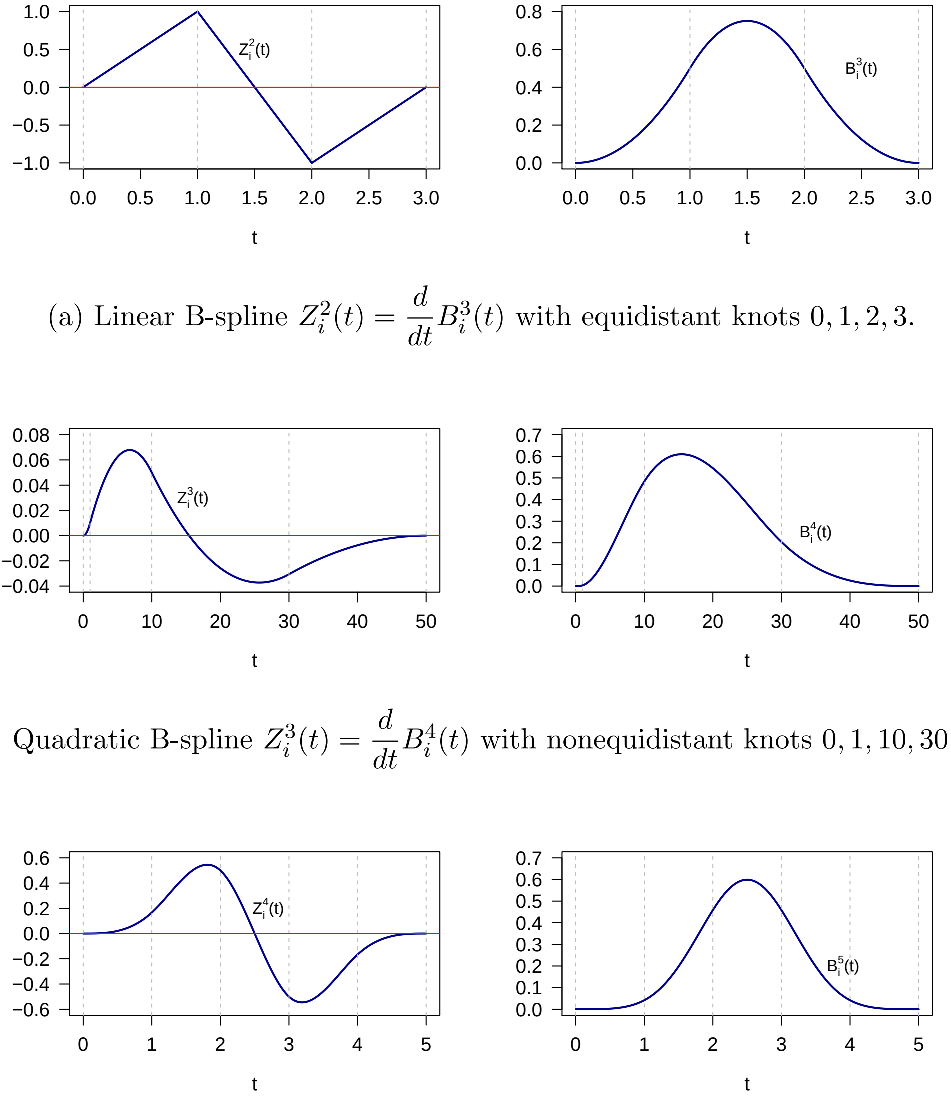
Now, regarding the definition (35), we are able to use spline functions $Z_i^{k+1}(t)$ which have zero integral on Ω (denoted as ZB-splines in the sequel). In the following, $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ denotes the vector space of polynomial splines of degree $k>0$, defined on a finite interval $\Omega=[a,b]$ with the sequence of knots $\Delta\lambda$ given in (21) and having zero integral on $[a,b]$, it means

$$\mathcal{Z}_k^{\Delta\lambda}[a,b] \;:=\; \left\{ s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b] \;:\; \int\limits_I s_k(t)\,dt = 0 \right\}. \tag{38}$$

Theorem 2.2 *The dimension of the vector space $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ defined by the formula (38) is $g+k$.*

Proof. For spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$, $s_k(t) = \sum\limits_{i=-k}^{g} b_i B_i^{k+1}(t)$ with the coincident additional knots it is known [8] that

$$\int\limits_{\Omega} s_k(t)\,dt \;=\; \frac{1}{k+1} \sum_{i=-k}^{g} b_i(\lambda_{i+k+1} - \lambda_i).$$



(a) Linear B-spline $Z_i^2(t) = \dfrac{d}{dt} B_i^3(t)$ with equidistant knots $0, 1, 2, 3$.

(b) Quadratic B-spline $Z_i^3(t) = \dfrac{d}{dt} B_i^4(t)$ with nonequidistant knots $0, 1, 10, 30, 50$.

(c) Cubic B-spline $Z_i^4(t) = \dfrac{d}{dt} B_i^5(t)$ with equidistant knots $0, 1, 2, 3, 4, 5$.

Figure 4: Example of piecewise polynomial functions $Z_i^{k+1}(t) := \frac{d}{dt} B_i^{k+2}(t)$ with $k \in \{1, 2, 3\}$. Vertical dashed gray lines indicate knot positions.

It means that B-spline coefficients of $s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}[a, b] \subset \mathcal{S}_k^{\Delta\lambda}[a, b]$ satisfy condition $0 = \sum_{i=-k}^{g} b_i(\lambda_{i+k+1} - \lambda_i) = \mathbf{Ab}$ with $\mathbf{A} = (\lambda_1 - \lambda_{-k}, \cdots, \lambda_{g+k+1} - \lambda_g)$,

$\mathbf{b} = (b_{-k}, \cdots, b_g)'$. And it is obvious that $\operatorname{codim}(\mathcal{Z}_k^{\Delta\lambda}[a,b]) = 1$, thus

$$\dim(\mathcal{Z}_k^{\Delta\lambda}[a,b]) \;=\; \dim(\mathcal{S}_k^{\Delta\lambda}[a,b]) - \operatorname{codim}(\mathcal{Z}_k^{\Delta\lambda}[a,b]) \;=\; g+k.$$

□

Theorem 2.3 *For the coincident additional knots (22), the functions $Z_{-k}^{k+1}(t)$, $\cdots, Z_{g-1}^{k+1}(t)$ form a basis for the space $\mathcal{Z}_k^{\Delta\lambda}[a,b]$.*

Proof. Since $M_i^{k+1}(t)$ form a basis for the spline space $\mathcal{S}_k^{\Delta\lambda}[a,b]$ and $Z_i^{k+1}(t) = M_i^{k+1}(t) - M_{i+1}^{k+1}(t)$, the functions $Z_i^{k+1}(t)$, $i = -k, \ldots, g-1$, are linearly independent and are elements of $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ with $\dim(\mathcal{Z}_k^{\Delta\lambda}[a,b]) = g+k$. Therefore $Z_i^{k+1}(t)$, $i = -k, \ldots, g-1$, form a basis in $\mathcal{Z}_k^{\Delta\lambda}[a,b]$. □

With regard to this theorem, each spline $s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}[a,b]$ has a unique representation

$$s_k(t) = \sum_{i=-k}^{g-1} b_i^z Z_i^{k+1}(t)\,. \tag{39}$$

Now we can proceed to a matrix notation of $s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}[a,b]$. With respect to (36) and (37), we are able to write the functions $Z_i^{k+1}(t)$ in matrix notation as

$$Z_i^{k+1}(t) = (k+1)\left(B_i^{k+1}(t)\,, B_{i+1}^{k+1}(t)\right) \begin{pmatrix} \dfrac{1}{\lambda_{i+k+1}-\lambda_i} & 0 \\ 0 & \dfrac{1}{\lambda_{i+k+2}-\lambda_{i+1}} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

that is, for ZB-spline basis of $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ we have

$$(Z_{-k}^{k+1}(t)\,, \ldots, Z_{g-1}^{k+1}(t)) = (B_{-k}^{k+1}(t)\,, \ldots, B_g^{k+1}(t))\mathbf{D}\mathbf{K}^z = \mathbf{B}_{k+1}(t)\mathbf{D}\mathbf{K}^z,$$

where matrix $\mathbf{D}$ is given in (31) and

$$\mathbf{K}^z = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix} \in \mathbb{R}^{g+k+1,g+k}. \tag{40}$$

It follows that each spline $s_k(t)$ from $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ can be written in the matrix notation using the standard B-spline basis with B-spline coefficients $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ fulfilling the condition (34) from Theorem 2.1 as

$$s_k(t) \;=\; \mathbf{Z}_{k+1}(t)\mathbf{b}^z \;=\; \mathbf{B}_{k+1}(t)\mathbf{D}\mathbf{K}^z\mathbf{b}^z = \mathbf{B}_{k+1}(t)\mathbf{b}, \tag{41}$$

with $\mathbf{Z}_{k+1}(t) = (Z_{-k}^{k+1}(t), \ldots, Z_{g-1}^{k+1}(t))$ and $\mathbf{b}^z = \left(b_{-k}^z, \ldots, b_{g-1}^z\right)'$. Note that the formula (41) provides a guideline how to convert the splines from $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ to splines with zero integral (with coefficients fulfilling (34)) from $\mathcal{S}_k^{\Delta\lambda}[a,b]$. This is particularly useful from the practical point of view as it allows to use existing codes in the statistical softwares for actual computations of the methods of FDA. For instance, the package `fda` of the statistical software R implements functional principal component analysis for a standard B-spline basis representation of functional data, hence it can be used for sampled (clr) density functions as in Section 5.

Example 2.2 *We consider knots* $\Delta\lambda := \lambda_0 = 0 = a < 2 < 5 < 9 < 14 < b = 20 = \lambda_5$. *The task is to find a cubic spline with the given sequence of knots and which has zero integral on the interval* $[0, 20]$. *It is evident that* $k = 3$, $g = 4$. *We consider the additional knots*

$$\lambda_{-3} = \lambda_{-2} = \lambda_{-1} = \lambda_0 = a = 0, \qquad 20 = b = \lambda_5 = \lambda_6 = \lambda_7 = \lambda_8.$$

The basis functions of the space $\mathcal{Z}_3^{\Delta\lambda}[0,20]$ *are plotted in Figure 5 (right). Every spline* $s_3(t) \in \mathcal{Z}_3^{\Delta\lambda}[0,20]$ *can be written as*

$$s_3(t) = \sum_{i=-3}^{3} b_i^z Z_i^4(t)\,. \tag{42}$$

Thus, e.g., for $\mathbf{b}^z = (b_{-3}^z, \ldots, b_3^z)' = (-1, -5, -15, -30, -30, 14, 5)'$ *the cubic spline function* $s_3(t)$ *with zero integral is plotted in Figure 5 (left).*

Unlike Approach I where the smoothing spline functions with zero integral were constructed upon standard B-spline basis system of functions $B_i^{k+1}(t)$, $i = -k, \ldots, g$, we can now use ZB-spline basis system of functions $Z_i^{k+1}(t)$, $i = -k, \ldots, g-1$ for this purpose. Accordingly, for an arbitrary $l \in \{1, \ldots, k-1\}$

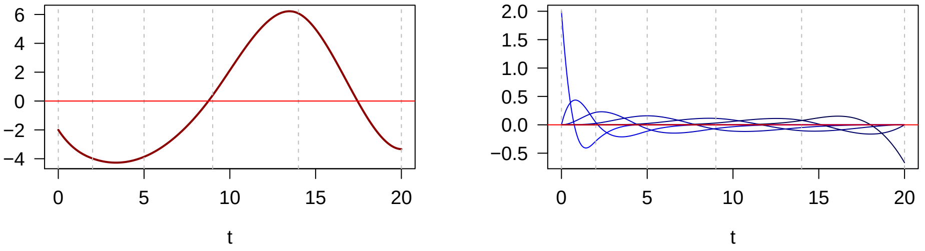


Figure 5: Cubic spline function $s_3(t)$ with the given coefficients $\mathbf{b}^z = (-1,-5,-15,-30,-30,14,5)'$ in $\mathcal{Z}_3^{\Delta\lambda}[0,20]$ (left) and ZB-spline basis system of $\mathcal{Z}_3^{\Delta\lambda}[0,20]$ (right). Vertical dashed gray lines indicate knot positions.

we aim to find a smoothing spline $s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}[a,b] \subset L_0^2([a,b])$ which minimizes the functional (24). The minimum is found for the spline $s_k^*(t)$ of the form (39) with ZB-spline coefficients $\mathbf{b}^{z*} = \left(b_{-k}^{z*}, \ldots, b_{g-1}^{z*}\right)'$ obtained as (see [24] for details)

$$\mathbf{b}^{z*} = \mathbf{V}^z \mathbf{z} \tag{43}$$

with

$$\mathbf{V}^z := \mathbf{G}^{-1}\mathbf{g}, \tag{44}$$

where

$$\mathbf{G} \; := \; (\mathbf{K}^z)'\mathbf{D}\left[(1-\alpha)\mathbf{S}_l'\mathbf{M}_{kl}\mathbf{S}_l + \alpha\mathbf{B}_{k+1}'(\mathbf{t})\mathbf{W}^{\mathbf{s}}\mathbf{B}_{k+1}(\mathbf{t})\right]\mathbf{D}\mathbf{K}^z \tag{45}$$

and

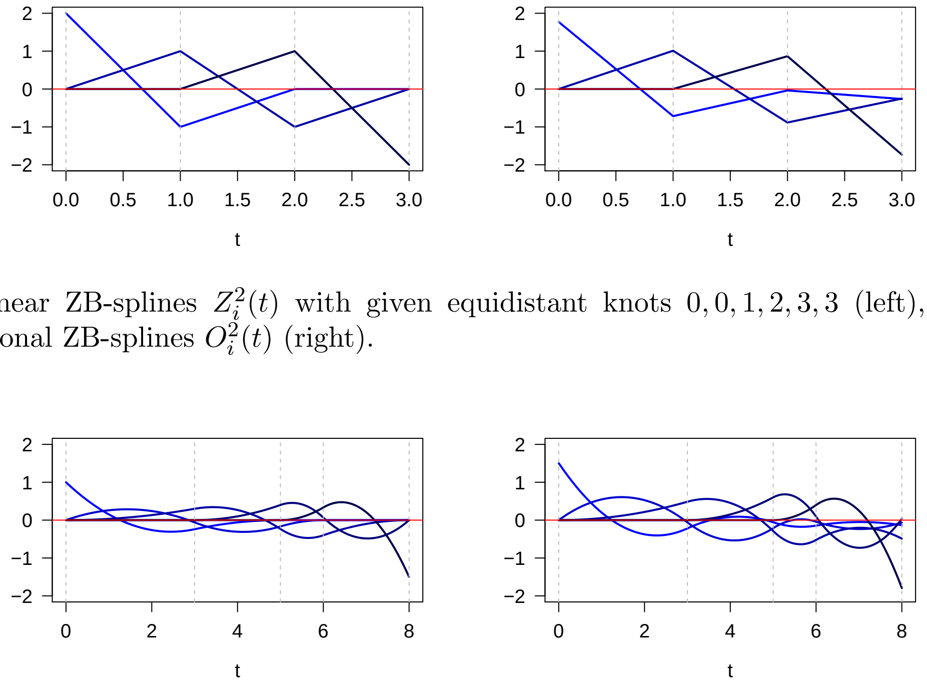
$$\mathbf{g} \; := \; \alpha(\mathbf{K}^z)'\mathbf{D}\mathbf{B}_{k+1}'(\mathbf{t})\mathbf{W}^s;$$

the matrices $\mathbf{B}_{k+1}(\mathbf{t}), \mathbf{D}, \mathbf{M}_{kl}, \mathbf{S}_l, \mathbf{K}^z$ are given in (30), (31), (32), (33), (40). Consequently, by considering the formula (41), the resulting smoothing spline in matrix notation using standard B-splines $B_i^{k+1}(x)$ is obtained as

$$s_k^*(t) \; = \; \mathbf{B}_{k+1}(t)\mathbf{D}\mathbf{K}^z\mathbf{b}^{z*},$$

where the vector $\mathbf{b}^{z*}$ is given in (43).

In some applications, the orthonormalization of the B-spline basis might be useful. Note that ZB-spline functions forming the basis system of $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ are by



(a) Linear ZB-splines $Z_i^2(t)$ with given equidistant knots $0, 0, 1, 2, 3, 3$ (left), linear orthogonal ZB-splines $O_i^2(t)$ (right).

(b) Quadratic ZB-splines $Z_i^3(t)$ with given nonequidistant knots $0, 0, 0, 3, 5, 6, 8, 8, 8$ (left), quadratic orthogonal ZB-splines $O_i^3(t)$ (right).

Figure 6: Orthogonalized ZB-spline basis system of $\mathcal{Z}_1^{\Delta\lambda}[0, 3]$ (first row) and $\mathcal{Z}_2^{\Delta\lambda}[0, 8]$ (second row). Vertical dashed gray lines indicate knot positions.

the default setting (36) non-orthogonal. The orthogonalized ZB-spline functions are obtained by using a linear transformation $\mathbf{\Phi}$ such that

$$\mathbf{\Phi}'\mathbf{\Phi} \;=\; \mathbf{\Sigma}^{-1},$$

where $\mathbf{\Sigma}$ represents the positive definite matrix

$$\mathbf{\Sigma} \;=\; \int\limits_a^b \mathbf{Z}_{k+1}(t)\mathbf{Z}'_{k+1}(t)\,\mathrm{d}t \;=\; \left(\int\limits_a^b Z_i^{k+1}(t)Z_j^{k+1}(t)\,\mathrm{d}t\right)_{i,j=-k}^{g-1}. \tag{46}$$

In a light of (41), the matrix $\mathbf{\Sigma}$ can be expressed as

$$\mathbf{\Sigma} \;=\; (\mathbf{K}^z)'\mathbf{D}\int_a^b \mathbf{B}'_{k+1}(t)\mathbf{B}_{k+1}(t)\,\mathrm{d}t\mathbf{D}\mathbf{K}^z = (\mathbf{K}^z)'\mathbf{D}\mathbf{M}_{k0}\mathbf{D}\mathbf{K}^z, \tag{47}$$

The linear transformation $\mathbf{\Phi}$ is not unique and can be computed for example by the Cholesky decomposition. The following basis functions

$$\mathbf{O}_{k+1}(t) = (O^{k+1}_{-k}(t), \ldots, O^{k+1}_{g-1}(t))' \tag{48}$$

obtained as

$$\mathbf{O}_{k+1}(t) = \mathbf{\Phi}\mathbf{Z}_{k+1}(t)$$

are orthogonal and have a zero integral. The linear and quadratic ZB-spline functions with zero integral and their ortogonalization are plotted in Figure 6. Consequently, the spline $s_k(t)$ with zero integral can be constructed as a linear combination of orthogonal ZB-splines with zero integral (48) in a form

$$s_k(t) \; = \; \sum_{i=-k}^{g-1} b^z_i O^{k+1}_i(t) \; = \; \mathbf{O}_{k+1}(t)\mathbf{b}^z;$$

or by considering the formula (41) in the form

$$s_k(t) \; = \; \mathbf{\Phi}\mathbf{B}_{k+1}(t)\mathbf{D}\mathbf{K}^z\mathbf{b}^z \tag{49}$$

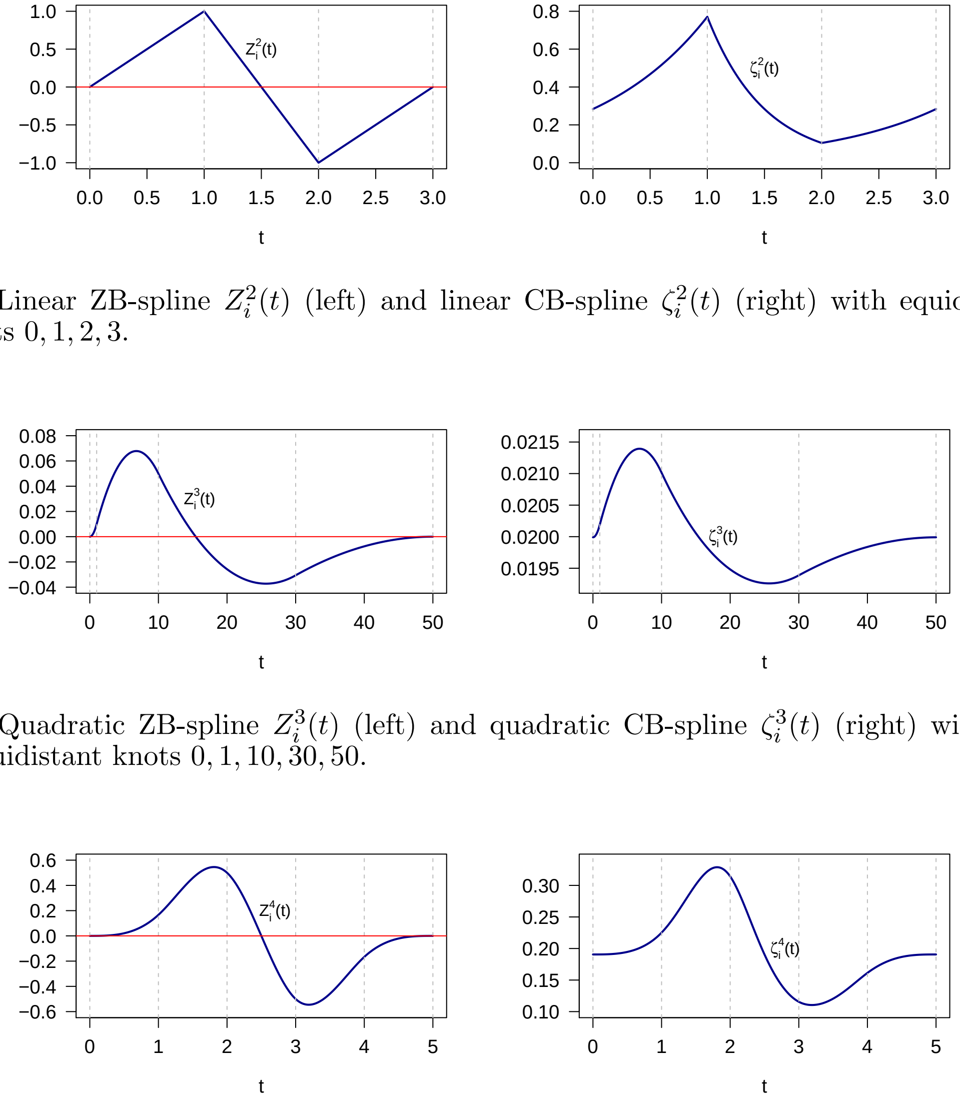
which is a preferable choice from the practical point of view.

Compositional splines in the Bayes spaces $\mathcal{B}^2(\lambda)$: Construction of spline functions directly in $L^2_0(\lambda)$ has important practical consequences, however, it is crucial also from the theoretical perspective. Expressing B-spline functions as functions in $L^2_0(\lambda)$ enables to transform them back to the original Bayes space $\mathcal{B}^2(\lambda)$ by using the exponential. It results in *compositional B-splines (CB-splines)*, obtained from (36) as

$$\zeta^{k+1}_i(t) =_{\mathcal{B}(\lambda)} \exp[Z^{k+1}_i](t), \quad i = -k, \ldots, g-1, \; k \geq 0. \tag{50}$$

Accordingly, for instance ZB-splines from Figure 4 can be now expressed directly in the Bayes space $\mathcal{B}^2(\lambda)$ as CB-splines, see Figure 7. As a consequence, it is immediate to define vector space $\mathcal{C}^{\Delta\lambda}_k[a,b]$ of compositional polynomial spline functions of degree $k > 0$, defined on a finite interval $\Omega = [a,b]$ with the sequence of knots $\Delta\lambda$. From isomorphism between $\mathcal{C}^{\Delta\lambda}_k[a,b]$ and $\mathcal{Z}^{\Delta\lambda}_k[a,b]$ it holds that

$$\dim\left(\mathcal{C}^{\Delta\lambda}_k[a,b]\right) = g + k.$$



(a) Linear ZB-spline $Z_i^2(t)$ (left) and linear CB-spline $\zeta_i^2(t)$ (right) with equidistant knots $0, 1, 2, 3$.

(b) Quadratic ZB-spline $Z_i^3(t)$ (left) and quadratic CB-spline $\zeta_i^3(t)$ (right) with nonequidistant knots $0, 1, 10, 30, 50$.

(c) Cubic ZB-spline $Z_i^4(t)$ (left) and cubic CB-spline $\zeta_i^4(t)$ (right) with equidistant knots $0, 1, 2, 3, 4, 5$.

Figure 7: Example of CB-spline functions of degree 1, 2 and 3 from the Figure 4. Vertical dashed gray lines indicate knot positions.

Moreover, from isometric properties of clr transformation (14) and (15) it follows that each compositional spline function $\xi_k(t) \in \mathcal{C}_k^{\Delta\lambda}[a, b]$ in $\mathcal{B}^2(\lambda)$ can be uniquely represented as

$$\xi_k(t) = \bigoplus_{i=-k}^{g-1} b_i^z \odot \zeta_i^{k+1}(t). \tag{51}$$

CB-splines $\zeta_i^{k+1}(t)$ forming the basis are by the default setting (36) not orthogonal. Their orthogonalization can be done in $L_0^2(\lambda)$ by using (49) and then back-transformed to $\mathcal{B}^2(\lambda)$.

The resulting compositional splines (with either orthogonal, or non-orthogonal CB-spline basis system) can be used for the representation of density functions directly in Bayes spaces. This is an important step in the construction of the FDA methods involving density functions. With CB-splines one has a guarantee that methods are developed consistently in the Bayes spaces. Moreover, the possibility of having an orthogonal basis enables to gain additional features resulting from orthogonality of finite dimensional projection in combination with approximate properties of spline functions.

Finally, compositional spline functions can be tuned according to a concrete problem, with the advantage of their direct formulation in the Bayes space sense.

2.2.3 Application: smoothing of Italian income data

To illustrate the smoothing procedures, Approach I and Approach II, respectively, we will use income data from the *Survey on Household Income and Wealth* (SHIW) conducted by the Italian Central Bank. They include almost 8000 interviewed households composed of 19907 individuals and 13266 income-earners and are freely available on the web [2]. We focus on annual net disposable income (composed of payroll income, i.e., net wages, salaries and fringe benefits, pensions and net transfers, net self-employment income and property income) of households in all $N = 20$ Italian regions, similarly as it was done in [23]. The regions were further grouped into three natural areas according to their geographical location to examine possible differences over regions (see Figure 8).

In the preprocessing step, the raw income data from individual regions were aggregated into histogram data as follows. The sampled values of incomes in each region were divided into $D = 9$ equally-spaced income classes determined by Sturges' rule [39] (i.e., its mean value over the regions was considered) for

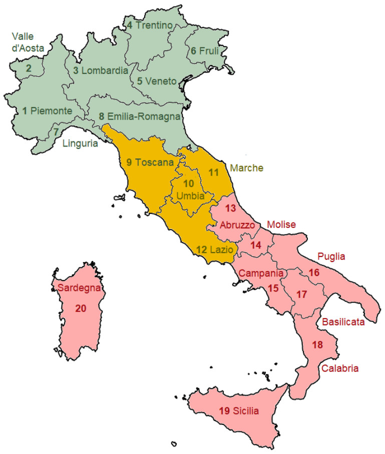


Figure 8: Map of Italy and its 20 regions with color distinguishing northern (green), middle (gold) and southern and island (red) regions according to the National statistical institute (ISTAT).

non-zero incomes up to 117.22 k€. Only incomes below the 99%-quantile were used and extreme values were excluded. Subsequently, the vectors of proportions of $D = 9$ income classes within each region $\mathbf{W}_i = (W_{iD}, \ldots, W_{iD})'$ were computed together with the raw discretized density data $\mathbf{y}_i = (y_{i1}, \ldots, y_{iD})', i = 1, \ldots, N$ which corresponds to the interval midpoints of income classes $\mathbf{t}_i = (t_{i1}, \ldots, t_{iD})', i = 1, \ldots, N$ (Tables 5 and 6 of Appendix B). These were obtained by dividing (not necessary normalized) proportions $\mathbf{W}_i, i = 1, \ldots, N$ by the length of the respective subintervals resulting from the partition of income interval $\Omega = [0, 117.22]$ k€ into income classes. The present zero-values were imputed by a model-based procedure [25]. Figure 9 shows an instance of two histograms together with raw data to be smoothed. To do so, their transformation into real vectors is conducted using clr transformation (17), resulting into vectors of raw clr transformed density data $\mathbf{z}_i = (z_{i1}, \ldots, z_{iD})'$ for $i = 1, \ldots, N$ (Table 7 of Appendix B). We note that as long as the histogram data are constructed on subintervals of the same length, i.e. with equally-spaced breakpoints, it enables to use the discrete clr transformation directly on the vector of proportions $\mathbf{W}_i, i = 1, \ldots, N$ by considering the scale invariance property; if not, the input of the clr transformation must be vectors with raw density data $\mathbf{y}_i, i = 1, \ldots, D$.

Having collected data $(t_{ij}, z_{ij}), i = 1, \ldots, N, j = 1, \ldots, D$, we aim to smooth them by (a) smoothing splines adapted for clr density functions using a standard

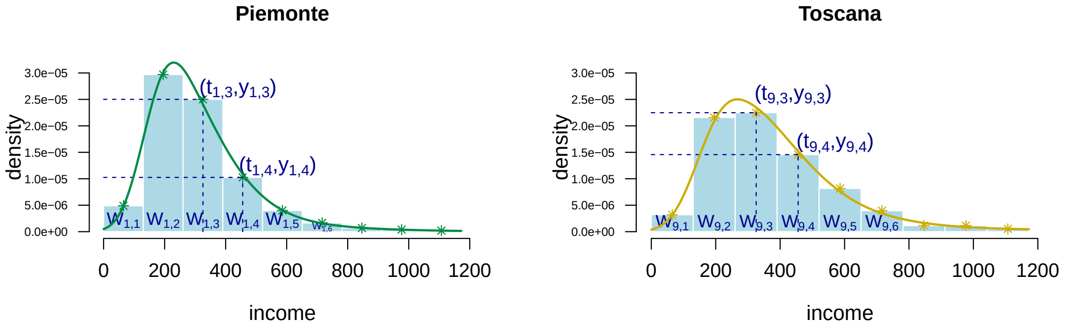


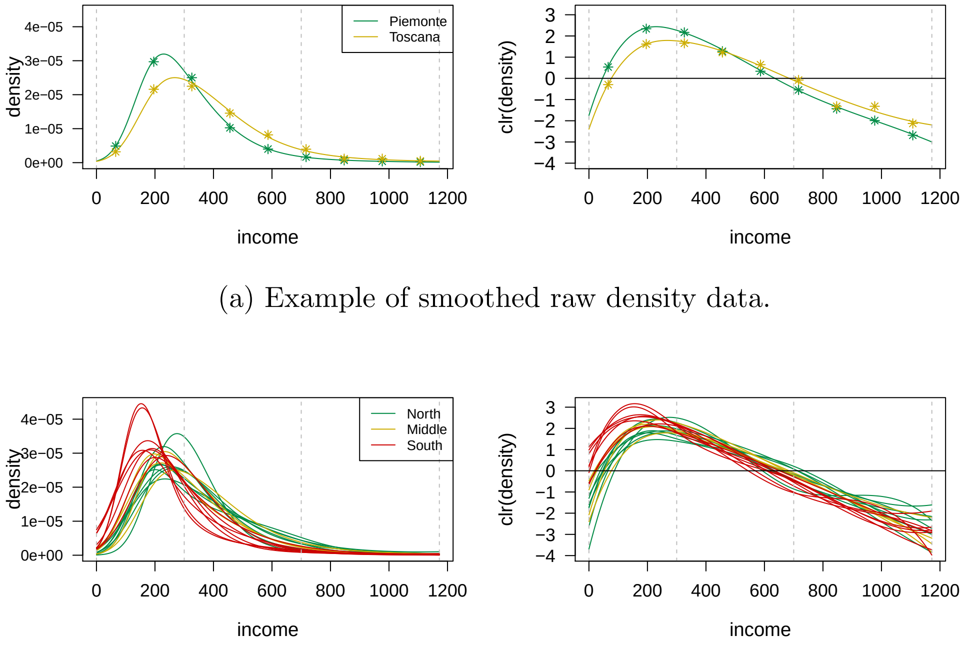
Figure 9: Histograms for Piemonte and Toscana region together with estimated probability density functions. Asterisks indicate discrete data $(t_{ij}, y_{ij}), i = 1, 9, j = 3, 4$ and $\mathbf{W}_i$ indicates indicate proportions of equidistant classes resulting from given partition of the interval $\Omega = [0, 117.22]$ k€; income is expressed in 10^3 k€.

B-spline basis system (Approach I) and (b) compositional smoothing splines (Approach II) using a system of ZB-spline basis functions from the $L_0^2(\lambda)$ space. The corresponding smoothing splines $s_k^i(t)$ on Ω which approximate density functions $f_i(t)$ for each $i = 1, \ldots, N$ are found by minimizing the functional (24) and having zero integral. For all N observations, the same strategy was followed to set the values of the input parameters for the smoothing procedures. Cubic smoothing splines were employed $(k = 3, l = 2)$ with four given knots at income values 0, 30, 70 and 117.22 k€, the vector of weights $\mathbf{w}_i^s$ for all input data was set to vectors of ones and the smoothing parameter α equals to 0.5 (i.e., the same importance is assigned to both smoothness of the resulting smoothing splines as well as to their approximative properties). Note that this choice corresponds to $\alpha = 1$ for the functional (26).

The resulting optimal smoothing splines and compositional smoothing splines represent the same approximations in a sense that they lead to the same functions $s_3^{i*}(t), i = 1, \ldots, N$, and are obtained via their clr representation as

$$s_3^{i*}(t) = \sum_{\nu=-3}^{2} b_{i,\nu}^* B_\nu^4(t) = \sum_{\nu=-3}^{1} b_{i,\nu}^{z*} Z_\nu^4(t), \quad i = 1, \ldots, N, \quad t \in \Omega; \tag{52}$$

the corresponding B-spline coefficients $\mathbf{b}_i^*$ and ZB-spline coefficients $\mathbf{b}_i^{z*}$ for $i = 1, \ldots, N$ are reported in Tables 1 and 2, respectively. Note that $\mathbf{b}_i^*, i = 1, \ldots, N$



(a) Example of smoothed raw density data.

(b) Smoothed raw density data.

Figure 10: Smoothed income density functions via smoothing splines in $\mathcal{B}^2(\lambda)$ (left) space and its clr transformation in $L_0^2(\lambda)$ (right) space with the same color resolution as in the map in Figure 8; income is expressed in 10^3 k€. Vertical dashed gray lines indicate knot positions.

can be derived directly from (41) using the output $\mathbf{b}_i^{z*}, i = 1, \ldots, N$. Nevertheless, only the Approach II enables to express explicitly the resulting splines in $\mathcal{B}^2(\lambda)$. Indeed, their representation with respect to CB-spline basis system of functions $\zeta_\nu^4(t) = \exp\left[Z_\nu^4\right](t), \nu = -3, \ldots, 1$ is given by

$$\xi_3^i(t) = \bigoplus_{\nu=-3}^{1} b_{i,\nu}^{z*} \odot \zeta_\nu^4(t), \quad i = 1, \ldots, N, \quad t \in \Omega.$$

An instance of two raw density data from Figure 9 is plotted together with smoothed curves in Figure 10a: in the $L_0^2(\lambda)$ space (right) and after the inverse transformation in the $\mathcal{B}^2(\lambda)$ space (left). The whole sample of smoothed density functions is displayed in Figure 10b; the color scheme matches those colors used for the geographical map (Figure 8). Visual inspection of Figure 10b suggests that a regional pattern may be present, as the northern regions seem to be associated

Region	loc.	B-spline coefficients, $\mathbf{b}_i^* = (b_{i,-3}^*, \ldots, b_{i,2}^*)'$, $i = 1, \ldots, N$					
Piemonte	N	-1.749	2.683	2.673	-1.057	-2.098	-3
Valle d'Aosta	N	-3.692	1.759	3.97	-2.909	-0.299	-2.346
Lombardia	N	-1.165	1.623	1.563	-0.174	-1.898	-1.609
Trentino	N	-2.62	2.544	1.392	0.515	-2.454	-2.308
Veneto	N	-0.565	1.674	2.512	-1.145	-1.819	-2.155
Friuli	N	-1.715	2.543	1.026	0.872	-2.42	-2.921
Liguria	N	-1.301	2.545	1.967	0.369	-2.651	-3.85
Emilia Romagna	N	-1.59	1.627	2.347	-0.797	-1.364	-2.729
Toscana	M	-2.396	1.612	2.266	-0.342	-1.865	-2.198
Umbria	M	-2.081	2.64	2.581	-0.552	-2.051	-3.837
Marche	M	-0.9	2.84	1.688	-0.154	-2.311	-3.179
Lazio	M	-0.473	2.299	2.204	-0.649	-1.908	-3.444
Abruzzo	S	-0.582	2.149	2.79	-1.497	-2.234	-1.898
Molise	S	-0.667	2.637	1.766	-0.361	-2.183	-2.941
Campania	S	0.176	4.575	1.122	-0.187	-2.972	-3.724
Puglia	S	0.982	3.099	2.489	-2.545	-0.597	-3.976
Basilicata	S	0.802	3.063	1.414	-0.176	-2.961	-2.653
Calabria	S	1.125	2.844	2.606	-2.069	-1.857	-2.833
Sicilia	S	-0.088	4.586	0.571	0.106	-3.049	-2.793
Sardegna	S	0.247	3.1	2.247	-0.641	-3.173	-2.88

Table 1: B-spline coefficients for optimal smoothing splines approximating income density functions of $N = 20$ Italian regions.

with higher incomes than the southern ones. This is probably related to the fact that a large number of businesses and industries are based in the north. The life cost is not homogeneous over the regions either, which may also play a major role in determining the actual salaries. Moreover, the highest variability seems to be present on boundaries of Ω and it is more pronounced on its left-hand side: low incomes dominate in the southern regions whereas in the central and the northern regions their incidence is significantly lower.

Once the functional data are reconstructed, we can proceed with the statistical analysis. The following section introduces functional regression analysis in the $\mathcal{B}^2(\lambda)$ space, the extension of SFPCA for the case of non-uniform reference measure is presented in Section 5 and applied to smoothed Italian income density functions.

Region	loc.	CB-spline coefficients, $\mathbf{b}_i^{z*} = (b_{i,-3}^{z*}, \ldots, b_{i,1}^{z*})', i = 1, \ldots, N$				
Piemonte	N	-13117.38	33827.37	112161.82	81173.56	35417.62
Valle d'Aosta	N	-27689.15	3101.96	119432.17	34199.26	27688.31
Lombardia	N	-8738.07	19656.85	65467.32	60374.37	18993.47
Trentino	N	-19650.11	24874.78	65671.57	80749.19	27243.95
Veneto	N	-4235.51	25058.29	98663.45	65116.43	25443.24
Friuli	N	-12866.1	31628.2	61707.04	87259.28	34484.86
Liguria	N	-9759.85	34783.47	92439.12	103264.55	45451.45
Emilia Romagna	N	-11926.89	16537.56	85324.37	61956.23	32212.77
Toscana	M	-17968.32	10238.6	76628.47	66598.39	25940.97
Umbria	M	-15606.04	30587.34	106209.41	90024.22	45295.48
Marche	M	-6753.12	42954.33	92429.05	87913.87	37528.47
Lazio	M	-3548.24	36686.45	101275.24	82268.39	40658.7
Abruzzo	S	-4366.94	33237.96	114990.47	71114.36	22410.71
Molise	S	-4999.18	41139.95	92894.57	82324.94	34722.56
Campania	S	1316.71	81370.66	114241.24	108770.21	43960.7
Puglia	S	7368.1	61603.74	134545.89	59960.23	46933.93
Basilicata	S	6011.31	59608.46	101045.52	95887.36	31320.53
Calabria	S	8436.74	58210.8	134570.72	73925.13	33439.39
Sicilia	S	-660.92	79602.63	96346.92	99445.84	32969.79
Sardegna	S	1855.1	56113.48	121968.24	103187.83	33996.89

Table 2: ZB-spline coefficients for compositional smoothing splines approximating income density functions of $N = 20$ Italian regions.

3 Statistical methods in unweighted Bayes spaces: Functional regression

Regression analysis is a key statistical tool to model a linear relationship between a response variable and a set of covariates. If the response or the predictors have functional nature, the functional regression analysis is to be considered. Although the general problem of functional regression has been extensively studied in the literature on FDA (i.e., for instance in [15, 36, 38]), a concise methodology for regression analysis in the presence of a distributional response has been proposed only recently in [41]. It aims to develop a general theoretical and computational setting allowing for the estimation and uncertainty assessment in linear models with a distributional response.

In this section, we firstly briefly recall the function-on-scalar regression model for data in L^2 spaces and subsequently, the main focus will be on a function-on-scalar model for a distributional response in Bayes spaces on closed interval $\Omega = [a, b]$. Similarly as in the L^2 setting, the key is to consider the B-spline representation of the PDF response observed as discrete (histogram) data. On these bases, the effective computational procedure is proposed and further discussed in this section.

3.1 Functional regression model in $L^2(\lambda)$

We here review the key notions on function-on-scalar regression that are deemed useful for our developments, by following [36, 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, $x_0 = 1$. Consider an N-dimensional vector of functional observations $\mathbf{y}(t)$ in $L^2(\lambda)$ on Ω, 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 $L^2(\lambda)$ on Ω. Call $\boldsymbol{\varepsilon}(t)$ an N-dimensional vector of i.i.d. (functional) random errors with zero-mean in $L^2(\lambda)$. The functional linear model for the i-th observation $y_i(t)$, $i = 1, ..., N$, associated with the regressors x_{ij}, $j = 0, ..., r$, is expressed as

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{r} x_{ij}\beta_j(t) + \varepsilon_i(t), \quad i = 1, \ldots, N, \quad t \in I, \tag{53}$$

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,

$$\mathrm{SSE}(\boldsymbol{\beta}) = \int_I [\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)]' \, [\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)] \, dt. \tag{54}$$

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

$$\mathrm{PENSSE}(\boldsymbol{\beta}) = \int_I [\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)]' \, [\mathbf{y}(t) - \mathbf{X}\boldsymbol{\beta}(t)] \, dt + \delta \int_I [L\boldsymbol{\beta}(s)]' \, [L\boldsymbol{\beta}(s)] \, ds, \tag{55}$$

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

Several computational methods have been proposed in the literature to minimize (54) or (55). In [36], 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

$$y_i(t) = \sum_{k=1}^{K_y} c_{ik}\varphi_k(t), \quad \beta_j(t) = \sum_{k=1}^{K_\beta} b_{jk}\psi_k(t), \tag{56}$$

in terms of known basis systems $\{\varphi_1, \ldots, \varphi_{K_y}\}$ and $\{\psi_1, \ldots, \psi_{K_\beta}\}$ (e.g., standard B-spline basis systems), with coefficients $\{c_{ik}\}$ and $\{b_{jk}\}$. Equivalently, we may express (56) 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 (56) 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 the estimation of the matrix of basis coefficients $\mathbf{B}$. They are found by minimizing the penalized least square fitting criterion, i.e.,

$$\begin{aligned}\text{PENSSE}(\boldsymbol{\beta}) = \int_I \left[\mathbf{C}\boldsymbol{\varphi}(t) - \mathbf{X}\mathbf{B}\boldsymbol{\varphi}(t)\right]' \left[\mathbf{C}\boldsymbol{\varphi}(t) - \mathbf{X}\mathbf{B}\boldsymbol{\varphi}(t)\right] dt \\ +\delta \int_I \left[L\mathbf{B}\boldsymbol{\varphi}(s)\right]' \left[L\mathbf{B}\boldsymbol{\varphi}(s)\right] ds. \qquad (57)\end{aligned}$$

Note that setting $\delta = 0$ yields the reformulation of (54) 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 (57) with respect to $\mathbf{B}$ it can be shown that the estimation of $\mathbf{B}$ is found as solution of the linear system

$$(\mathbf{X}'\mathbf{X}\mathbf{B}\mathbf{Q} + \delta\mathbf{B}\mathbf{P}) = \mathbf{X}'\mathbf{C}\mathbf{Q}. \qquad (58)$$

System (58) can be equivalently reformulated using the Kronecker product $\otimes$ as

$$\left[\mathbf{Q} \otimes (\mathbf{X}'\mathbf{X}) + \mathbf{P} \otimes \delta\mathbf{I}\right] \text{vec}(\mathbf{B}) = \text{vec}\left(\mathbf{X}'\mathbf{C}\mathbf{Q}\right). \qquad (59)$$

Matrix $\mathbf{B}$ is thus obtained as solution of a system of linear equations of dimension $p \times K$ and the resulting estimations of regression parameters as $\widehat{\boldsymbol{\beta}}(t) = \widehat{\mathbf{B}}\boldsymbol{\varphi}(t), t \in \Omega$.

3.2 Functional regression model in $\mathcal{B}^2(\lambda)$

In this subsection, a function-on-scalar regression model in $\mathcal{B}^2(\lambda)$ is introduced as a counterpart of the model (53). We assume the dependent variable $y(t), t \in \Omega$ to be an element of $\mathcal{B}^2(\lambda)$ 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 $\mathcal{B}^2(\lambda)$ of the form

$$y_i(t) = \beta_0(t) \oplus \bigoplus_{j=1}^{r} \left[x_{ij} \odot \beta_j\right](t) \oplus \varepsilon_i(t) \qquad (60)$$

where ε_i denotes a zero-mean functional error (or residual) in $\mathcal{B}^2(\lambda)$, $i = 1, \ldots, N$, and the unknown functions β_j, $j = 0, ..., r$, belong to $\mathcal{B}^2(\lambda)$ 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 $\mathcal{B}^2(\lambda)$

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\varepsilon_i\|^2_{\mathcal{B}^2(\lambda)} = \sum_{i=1}^{N} \left\| \bigoplus_{j=0}^{r} [x_{ij} \odot \beta_j] \ominus y_i \right\|^2_{\mathcal{B}^2(\lambda)} . \tag{61}$$

Note that (61) is the counterpart of SSE (54) in the Bayes Hilbert space $\mathcal{B}^2(\lambda)$; in fact, it also represents the analogue of compositional SSE in $\mathcal{B}^2(\mathsf{P}^c)$ [11] in infinite dimensions. Applying the clr transformation (14) to both sides of the model (60) yields

$$\mathrm{clr}_\lambda(y_i)(t) = \mathrm{clr}_\lambda(\beta_0)(t) + \sum_{j=1}^{r} \left[x_{ij} \cdot \mathrm{clr}_\lambda(\beta_j)\right](t) + \mathrm{clr}_\lambda(\varepsilon_i)(t), \quad i = 1, \ldots, N, \tag{62}$$

that enables one to reformulate the objective SSE (61) equivalently in the L^2 sense as

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\mathrm{clr}_\lambda(\varepsilon_i)\|^2_{L^2(\lambda)} = \sum_{i=1}^{N} \left\| \sum_{j=0}^{r} [x_{ij} \cdot \mathrm{clr}_\lambda(\beta_j)] - \mathrm{clr}_\lambda(y_i) \right\|^2_{L^2(\lambda)} . \tag{63}$$

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

As a next step, since both the observed functions $\mathrm{clr}_\lambda(y_i)(t)$, $i = 1, \ldots, N$, and regression parameters $\mathrm{clr}_\lambda(\beta_j)(t)$, $j = 0, \ldots, r$, are elements of $L^2_0(\lambda)$, their basis expansion fulfilling the zero-integral constraint on Ω using a given basis system $\{\varphi_k, k = 1, ..., K\}$ must be considered, i.e.,

$$\int_I \mathrm{clr}_\lambda(y_i(t))dt = \int_I \sum_{k=1}^{K} c_{ik}\varphi_k(t)dt = 0; \quad \int_I \mathrm{clr}_\lambda(\beta_j(t))dt = \int_I \sum_{k=1}^{K} b_{jk}\varphi_k(t)dt = 0. \tag{64}$$

Both approaches to basis expansions designed for densities outlined in Sections 2.2.1 and 2.2.2, respectively, can be used when estimating the linear model (60).

Nevertheless, we note that, regarding to Theorem 2.1, the former leads to constraints on the coefficients $\{c_{ik}\}$, $\{b_{jk}\}$ and consequently on model singularities. Although this can be overcome by using the latter method based on compositional splines, both of them in fact lead to the same estimations of regression parameters. Therefore, in the following we will mainly focus on the consequences related to the former approach as developed in [41].

3.2.1 Regression modeling of B-spline coefficients using Approach I

Let us consider the B-spline representations for the clr transformed observations of the response density, i.e., $\mathrm{clr}_\lambda(y_i)(t), i = 1, \dots, N$, of the form

$$s_k^i\left(t\right) = \sum_{j=-k}^{g} Y_{i,j+k+1} B_j^{k+1}\left(t\right), \tag{65}$$

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

$$\mathbf{Y}_{(i)} = \mathbf{V}\mathbf{z}_{(i)}, \quad i = 1, \dots, N; \tag{66}$$

the matrix $\mathbf{V}$ of dimensions $(g+k+1) \times D$ is given in (29) and $\mathbf{z}_{(i)} = (z_{i1}, \dots, z_{iD})'$, $i = 1, \dots, N$ are vectors of clr transformed raw density data. If the same B-spline basis system is used for all the data, (66) can be expressed in matrix notation as

$$\underline{\mathbf{Y}} \;=\; \underline{\mathbf{Z}}\mathbf{V}', \tag{67}$$

where $\underline{\mathbf{Y}}$, $\underline{\mathbf{Z}}$ are the matrices of dimensions $N \times (g+k+1)$ and $N \times D$, respectively, having the following form,

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

Consequently, we can express the model (60) in the form of a multivariate regression model. Following the given notation, spline coefficients for the i-th observation $y_i(t)$ are denoted by $\mathbf{Y}_{(i)} = (Y_{i,1}, \dots, Y_{i,g+k+1})'$, $i = 1, 2, \dots, N$, and 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 (60) the multivariate linear regression model of the form

$$\underline{\mathbf{Y}}_{(N\times(g+k+1))} = \mathbf{X}_{(N\times p)}\mathbf{B}_{(p\times(g+k+1))} + \underline{\boldsymbol{\varepsilon}}_{(N\times(g+k+1))}, \tag{68}$$

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

$$\widehat{\mathbf{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{g+k+1}), \tag{69}$$

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

$$\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}_j,\ j = 1, \ldots, g+k+1. \tag{70}$$

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., $\mathcal{M}(\mathbf{X}) = \{\mathbf{X}\mathbf{u} : \mathbf{u} \in \mathbb{R}^p\}$.

Because the realization of the 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

$Y_{i,1}, \dots, Y_{i,g+k+1}$ are linearly dependent. Indeed, one has that

$$\sum_{j=1}^{g+k+1} Y_{ij}(\lambda_j - \lambda_{j-k-1}) = 0, \tag{71}$$

due to Theorem 2.1. Accordingly, one may expect that a similar constraint applies to the corresponding estimated regression coefficients, as stated by the following result.

Proposition 3.2.1 *If* $\sum_{j=1}^{g+k+1} Y_{ij}(\lambda_j - \lambda_{j-k-1}) = 0$ *for all* $i = 1 \dots, N$*, then* $\sum_{j=1}^{g+k+1} \widehat{\beta}_{sj}(\lambda_j - \lambda_{j-k-1}) = 0$ *for all* $s = 0, \dots, r$*.*

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

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

□

Note that whenever the same B-spline basis is employed for all the observations of the response – as it is usually the case – the latter constraint from Preposition 3.2.1 introduces a singularity into the regression model (68), which may affect parameter inference. Similarly as in multivariate regression [17], 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. Since vectors $\mathbf{Y}_{(i)}$, $i = 1, ..., N$, form a hyperplane $\mathcal{H}$ of dimension $g + k$, orthogonal to the normal vector $(\lambda_1 - \lambda_{-k}, \dots, \lambda_{g+k+1} - \lambda_g)'$ 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 (64) – 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, that in turn correspond to the Simplicial Functional Principal Components (SFPCs) of the smoothed densities $y_1(t), ..., y_N(t)$ [21]. However, note that the BLUE estimation (69) 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. Of course, the singularity problem can be prevented by considering ZB-spline basis system from $L_0^2(\lambda)$, so that the response is expressed through a set of unconstrained coefficients, namely ZB-spline coefficients.

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 representing histograms). The number of knots and the number of classes on Ω upon which $\mathbf{W}_i$ are built are indeed related, as the former cannot exceed the latter. The problem of setting the discretization on Ω and the number of knots is affected by 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 Sturges' rule [39], which has been already used in the example of Section 2.2.3. 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 dealing with metabolomics data presented further in this section.

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 from Section 3.2 is applied (hereafter named “regression-smoothing”), and (b) a compositional regression [11] is applied, estimating the model

$$\mathbf{z}_i \;=\; \boldsymbol{\beta}_0^{(Z)} + \sum_{j=1}^{r} \boldsymbol{\beta}_j^{(Z)} x_{ij} + \boldsymbol{\epsilon}_i, \tag{72}$$

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 presented here and that one proposed in [11]

$$\begin{tikzcd}
\underline{\mathbf{Z}} \arrow[r, "smoothing"] \arrow[d, "regression"'] & \underline{\mathbf{Y}} \arrow[d, "regression"] \\
\widehat{\underline{\mathbf{Z}}} \arrow[r, "smoothing"'] & \widehat{\underline{\mathbf{Y}}}
\end{tikzcd} \tag{73}$$

From (69), the matrix of predicted coefficients $\underline{\mathbf{Y}}$ is obtained as

$$\widehat{\underline{\mathbf{Y}}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Y}}, \tag{74}$$

while for the model (72) one has

$$\widehat{\underline{\mathbf{Z}}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Z}}. \tag{75}$$

Plugging-in (67) in (74) 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 $\widehat{\mathbf{z}}_i$, $i = 1, \dots, N$, are considered, the matrix of the corresponding B-spline coefficients is obtained as

$$\widehat{\widehat{\underline{\mathbf{Y}}}} \;=\; \widehat{\underline{\mathbf{Z}}}\mathbf{V}'_Z. \tag{76}$$

In order to guarantee that $\mathbf{V}_Z$ coincides with the matrix $\mathbf{V}$ in (67), 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 (75), 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., $\underline{\widehat{\widehat{\mathbf{Y}}}} = \underline{\widehat{\mathbf{Y}}}$. 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, introduced in Section 3.1.

It should be noted that, although under particular conditions the "smoothing-regression" and "regression-smoothing" approaches are equivalent, the proposed framework provides a much 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 to the same set of classes, which may not be the optimal one 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, [29, 30, 31]). Dealing with high-dimensional (compositional) data from a discrete viewpoint may yield issues related to the curse of dimensionality, which are completely overcome with a functional viewpoint.

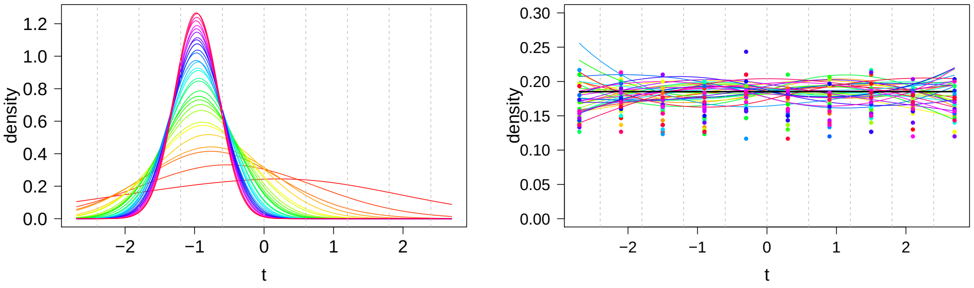
3.3 Simulation study

3.3.1 Assessing the effects of smoothing on regression

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

$$y_i(t) = \beta_0(t) \oplus [x_i \odot \beta_1](t) \oplus \varepsilon_i(t), \quad t \in \Omega = [-3, 3], \quad i = 1, \ldots, 30, \tag{77}$$

where ε is the random error, whose mean is the neutral element of perturbation in $\mathcal{B}^2(\lambda)$, i.e., the uniform distribution. Specifically, for each observation i, 500 realizations are generated from a uniform distribution on $\Omega = [-3, 3]$ and then smoothed via optimal smoothing splines (Approach I) to represent the errors ε_i through a B-spline basis. The smoothing procedure is designed to reproduce the estimation strategy which will be applied in the case study, and is based on qua-



(a) Reference response.

(b) Smoothed errors via B-splines.

Figure 11: Example of starting data of the first iteration – simulated (true) density response (left) with model errors (right) represented by optimal smoothing splines (Approach I; Section 2.2.1). Vertical dashed grey lines indicate partition of interval Ω with 10 classes.

dratic splines ($k = 2, l = 1$), with smoothing parameter $\alpha = 0.99$ and five equally spaced knots. The regression parameters are set to truncated Gaussian densities $N(\mu_i, \sigma_i^2)$, $i = 0, 1$, with support on Ω. For the intercept β_0 the parameters are 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 shall be estimated based on three different parameter settings, one determined by using the Sturges' rule, one above and one below the first. Specifically, with the previous model settings, Sturges' rule suggests an optimal value of $D = 10$ classes. The two additional values considered are thus $D = 7$ and $D = 14$. The simulated functional dataset is computed for each of such settings taking the model (77) for 30 equally spaced values x_i in the interval $[0.01, 10]$. To assess the performances of the method in estimating the parameters, the simulation is repeated $K = 30$ times. Figure 11 represents the observed response for the first iteration, together 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 $= \|\beta_{l_k} \ominus \hat{\beta}_{l_k}\|^2_{\mathcal{B}^2(\lambda)}$, $l = 0, 1$, $k = 1, ..., 30$, was considered. The top panels of Figure 12 display boxplots of the integrated square errors for the number of classes $D = 7, 10, 14$. Simulations

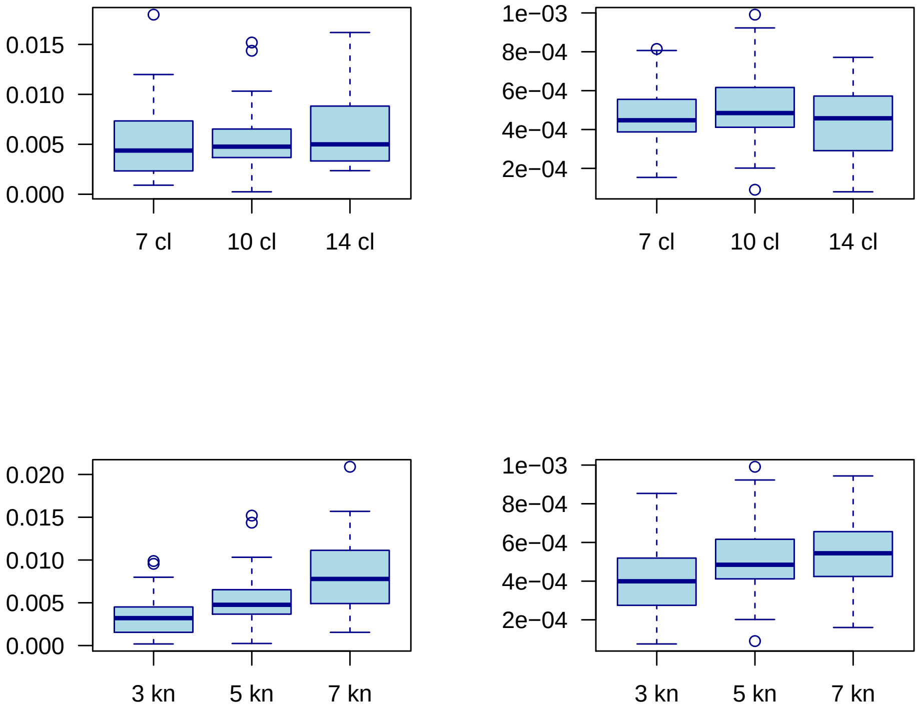


Figure 12: Boxplots of ISE between the true density parameter functions β_0 (left) and β_1 (right) and their estimates obtained by iteration procedure from the simulated model (77) – Model 500. Top panels: sensitivity to number of classes (cl.): 7, 10 (Sturges' rule) and 14. Bottom panels: sensitivity to number of knots in $\{3, 5, 7\}$.

shows that the Sturges' rule can be considered as a reasonable choice as the estimates of both parameters do not appear to be sensitive to that parameter setting.

Having fixed the number of classes according to Sturges' rule, the 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 12 seem to suggest the use of a moderate number of knots, as a 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, [21]), compatible with the low number of knots suggested by

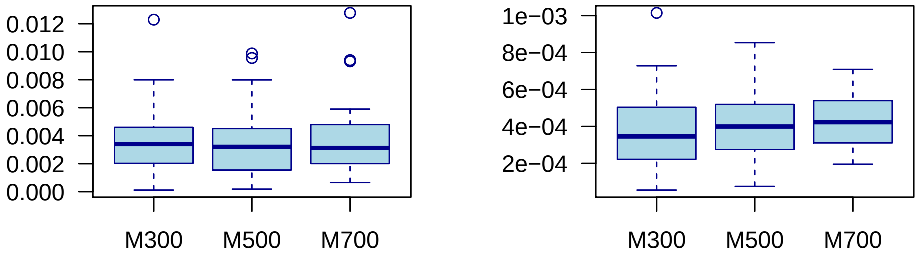


Figure 13: Comparison of boxplots of ISE for models with different number of starting data with optimal parameter setting – Model 300 (9 classes, 5 knots), Model 500 (10 classes, 3 knots), Model 700 (10 classes, 5 knots) – for β_0 left and β_1 right.

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 suggests 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 13). Note that all the simulation settings here considered are based upon a relatively high number of data, which is the setting for which the method is proposed. Indeed, the main issue with the sample size is related with the need of estimating the entire response distribution from the data 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 [11].

3.3.2 Comparison of the Bayes approach with competitors in $L^2(\lambda)$

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 11) and the same scalar regressors $\{x_i, i = 1, ..., 30\}$ as in Section

3.3.1. To this end, the following models are considered: (a) a function-on-scalar model in $L^2(\lambda)$

$$y_i(t) = \alpha_0(t) + \alpha_1(t)x_i + \zeta_i(t), \tag{78}$$

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

$$\log(y_i(t)) = \gamma_0(t) + \gamma_1(t)x_i + \eta_i(t), \tag{79}$$

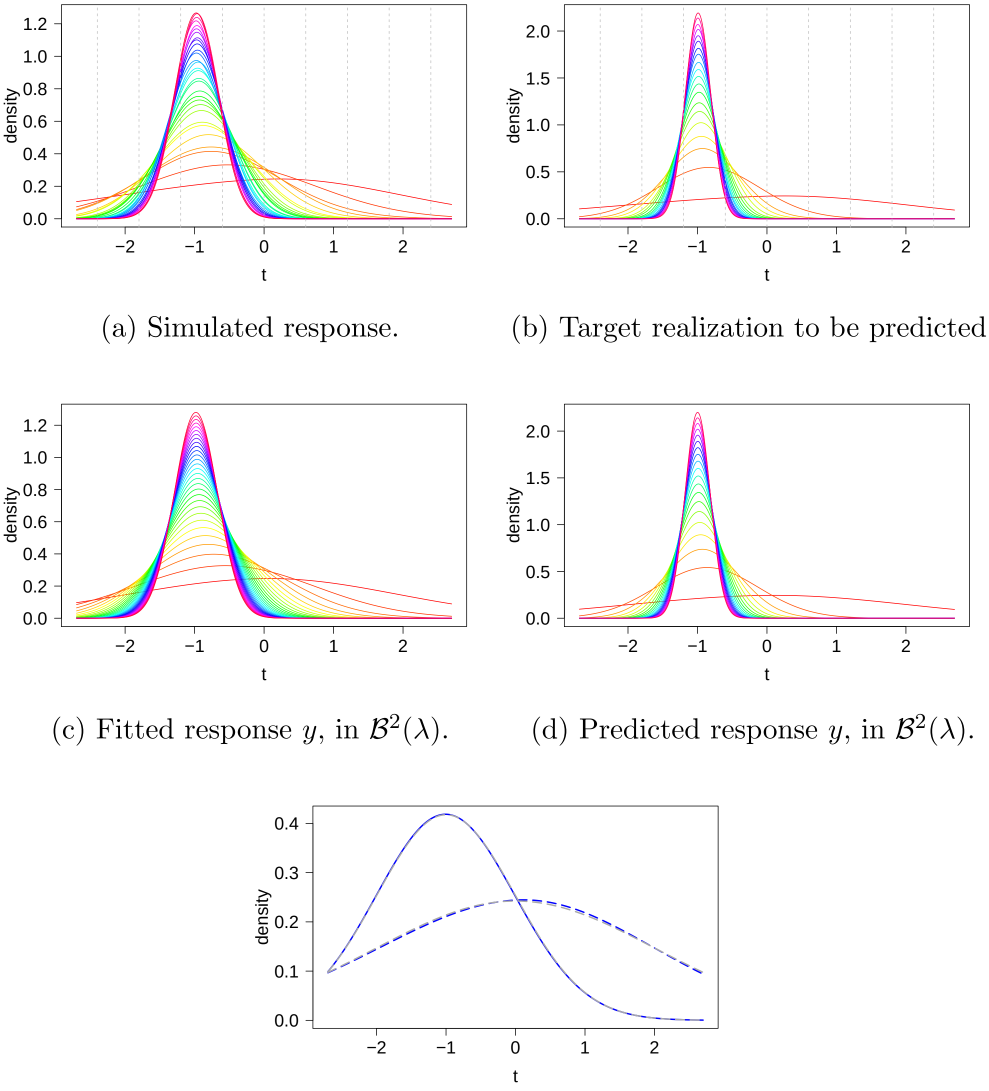
with η_i random errors in $L^2(\lambda)$, with mean zero. Note that using a logarithmic transformation is very common for data on a relative scale, and preserves positiveness, but does not guarantee that the resulting estimates keep scale invariance with the possibility of a unit integral representation. Estimation of the models (78) and (79) is 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 shall be compared in terms of (i) goodness of fit on the (simulated) response in Figure 14a and (ii) quality of predictions in correspondence of 20 equally spaced new values of the regressors x in $[0, 30]$ (Figure 14b).

Figure 14 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 14e).

In fact, very different results are obtained when using the geometry of $L^2(\lambda)$ (Figure 15). The model in $L^2(\lambda)$ clearly provides poorer estimations since fitted (Figure 15a) as well as predicted responses (Figure 15b) do not follow the integral constraint and exhibit negative values. Moreover, the difference between predicted curves in $L^2(\lambda)$ and in $\mathcal{B}^2(\lambda)$ 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 a L^2 approach (see Figure 16), as the resulting densities are guaranteed to be positive. However, fitted and predicted



(a) Simulated response.

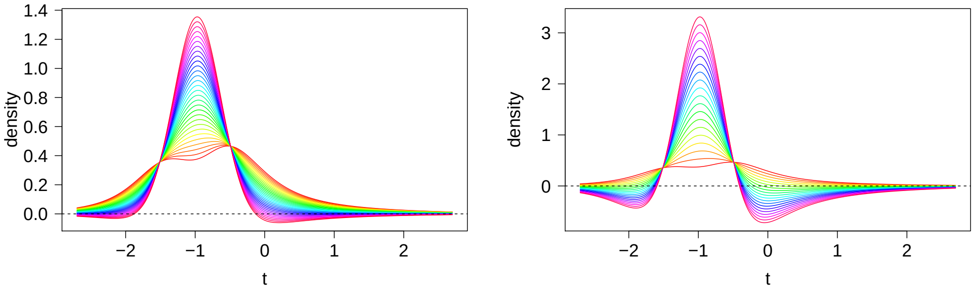
(b) Target realization to be predicted.

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

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

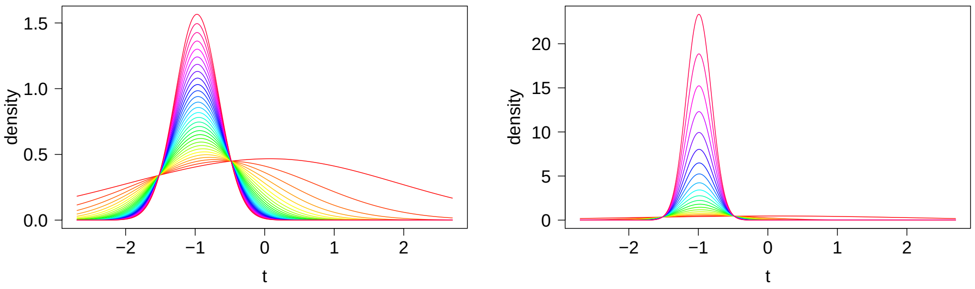
(e) True (grey line) and estimated (blue line) parameters.

Figure 14: Fitted response and estimated parameters in $\mathcal{B}^2(\lambda)$ space. Vertical dashed grey lines indicate partition of interval Ω with 10 classes.



(a) Fitted response y, in $L^2(\lambda)$. (b) Predicted response y, in $L^2(\lambda)$.

Figure 15: Fitted and predicted model in $L^2(\lambda)$ space.



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

Figure 16: Fitted and predicted model in $L^2(\lambda)$ space using log-transformation.

responses do not honor the unit integral constraint, thus providing unsatisfactory results.

3.4 Application: 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, and were not yet used elsewhere. Although they are not publicly available, their aggregations are given in Appendix C. 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.

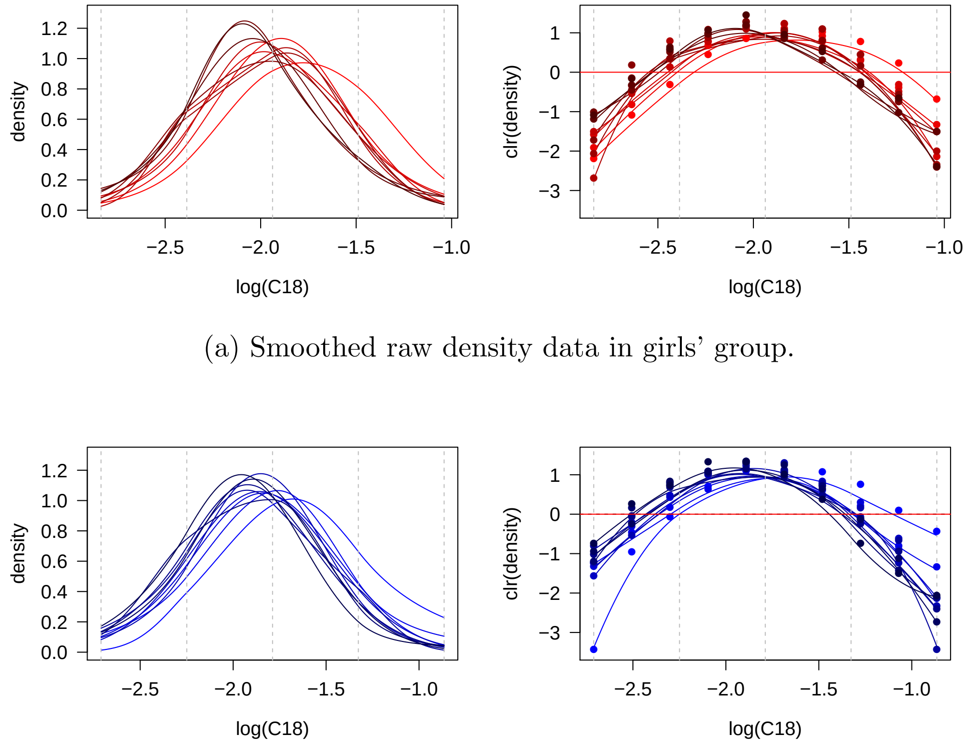
For the purpose of modeling the dependence of C18 distribution on weight through functional regression models, the C18 distribution was assessed from sampled data as follows. The values of the logarithm of C18 were divided into 10 groups of equal size according to the logarithm of weight, and represented 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 $N = 10$ groups, the distribution of log(C18) was estimated empirically, by dividing in equally spaced classes and computing the frequency within each class. The number of optimal classes were computed as previously based on Sturges' rule, resulting in mean value 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 $\Omega_g = [-2.936, -0.939]$ and $\Omega_b = [-2.813, -0.763]$. Tables 8 and 9 in Appendix C list the vectors of proportions $\mathbf{W}_i = (W_{i1}, \ldots, W_{iD})'$, $i = 1, \ldots, D$, of log(C18) within each group of weights, together with the midpoints of the classes $t_j, j = 1, \ldots, 10$ (index i is omitted for the midpoints values as they coincide in all $N = 10$ weight groups). In this setting, since the vectors of proportions – histogram data – are constructed upon the same partition of the intervals, Ω_g and Ω_b, respectively, with equally-spaced breakpoints, they can be directly used in order to get real vectors of clr transforms of raw density functions for all $N = 10$ groups, for both girls and boys, respectively. It results in the vectors $\mathbf{z}_i = (z_{i1}, \ldots, z_{iD})'$, $i = 1, \ldots, D$, reported in Tables 10 and 11 of Appendix C.

As a second step of the analysis, the raw observations (t_j, z_{ij}) within the $N = 10$ weight groups were turned into the smooth density functions by using smoothing splines (Approach I; Section 2.2.1) with support Ω_g and Ω_b, for girls and boys. In both cases (i.e., for girls and boys) the same strategy was followed to set the values of the input parameters for smoothing procedure. 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 equally spaced knots in the set $\{3, 5, 7\}$. The optimal smoothing spline $s_k(t)$ on Ω was found as to minimize the penalized functional (24) with respect to condition (25) where parameter α was set to $\alpha = 0.99$ in order to be as close as possible to input data, and the weights were set to $w_j^s = 1$, for $j = 1, \ldots, D$. Note that the same approximations would be obtained via compositional smoothing splines (Approach II; Section 2.2.2) if the setting of input parameters of the smoothing procedure is preserved. The resulting smoothed clr density functions $\mathrm{clr}_\lambda(y_i)(t) \in L_0^2(\lambda)$, $i = 1, ..., N$, are displayed in Figure 17 together with the corresponding density functions $y_i(t) \in \mathcal{B}^2(\lambda)$, $i = 1, ..., N$, obtained by applying the inverse clr transformation to the smoothed clr functions, i.e., $y_i(t) = \exp\left[\mathrm{clr}_\lambda(y_i)\right](t), i = 1, \ldots, N$. Data are plotted on red (in girls' group) and blue scale (in boys' group) distinguishing the weight groups – a low intensity of the colors is associated with a lower weight of newborns, while its large intensity with a higher weight of newborns.

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 $\mathcal{B}^2(\lambda)$,

$$y_i(t) = \beta_0(t) \oplus \left[\log(w_i) \odot \beta_1\right](t) \oplus \varepsilon_i(t), \quad i = 1, \ldots, 10. \tag{80}$$

By considering the same standard B-spline basis functions $B_{-2}^3(t), \ldots, B_3^3(t)$ from $L^2(\lambda)$ for the response $\mathrm{clr}_\lambda(y)(t)$, the regression parameters $\mathrm{clr}_\lambda(\beta_0)(t)$, $\mathrm{clr}_\lambda(\beta_1)(t)$ and the error $\mathrm{clr}_\lambda(\varepsilon)(t)$, model (80) can be written as a multivariate model for the B-spline coefficients $\mathbf{Y}_{(i)} = (Y_{i,1}, \ldots, Y_{i,6})'$, in matrix form as $\underline{\mathbf{Y}} = \mathbf{X}\mathbf{B} + \underline{\boldsymbol{\varepsilon}}$. 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 3, together with the estimates of their standard deviations. The corresponding estimates of the regression functions $\mathrm{clr}_\lambda(\beta_0)(t)$ and $\mathrm{clr}_\lambda(\beta_1)(t)$ are displayed in Figure 18, together with their counterparts in



(a) Smoothed raw density data in girls' group.

(b) Smoothed raw density data in in boys' group.

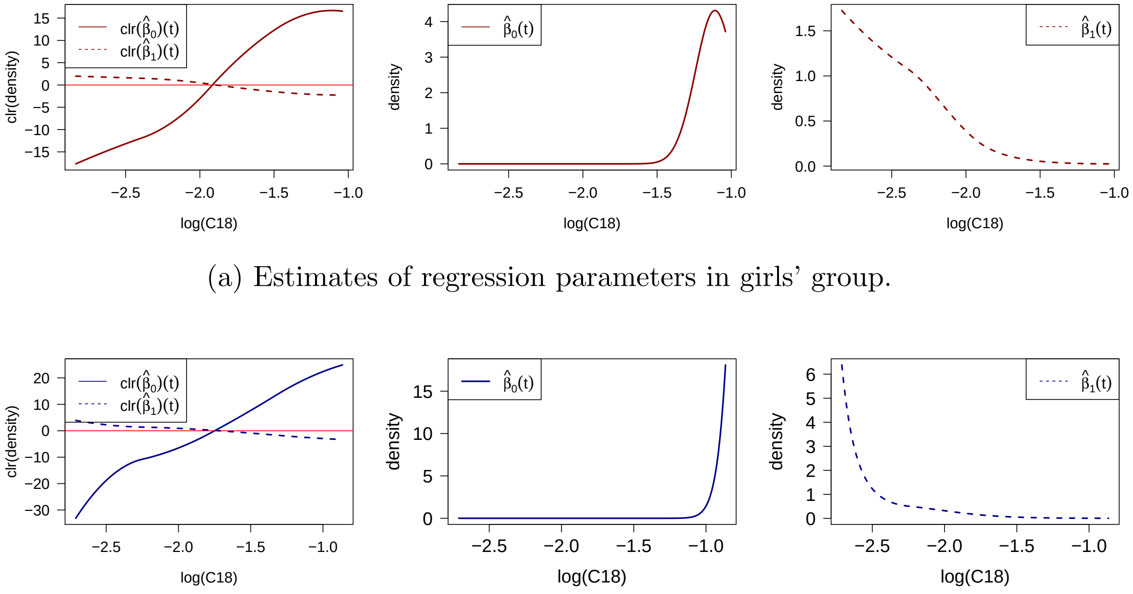
Figure 17: Smoothed density functions of log concentrations of metabolite C18 (log(C18)) via optimal smoothing splines (Approach I; Section 2.2.1) in $\mathcal{B}^2(\lambda)$ (left) space and its clr transformation in $L_0^2(\lambda)$ (right) spaces: top panels for girls, bottom panels for boys. Vertical dashed gray lines indicate knot positions.

$\mathcal{B}^2(\lambda)$ with the same color resolution, i.e., 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 18 (top panels). 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 the newborn is 1 gram. Nevertheless, the coefficient $\beta_0(t)$ acts as a shift in the model – in sense of the geometry of Bayes spaces – towards a density highly concentrated in the right tail of domain Ω_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 the domain Ω_g, and

	Estimates of regression parameters $\boldsymbol{\beta}_{\cdot} = (\beta_{\cdot 1}, \dots, \beta_{\cdot 6})'$					
$\widehat{\boldsymbol{\beta}}_0^g$	-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

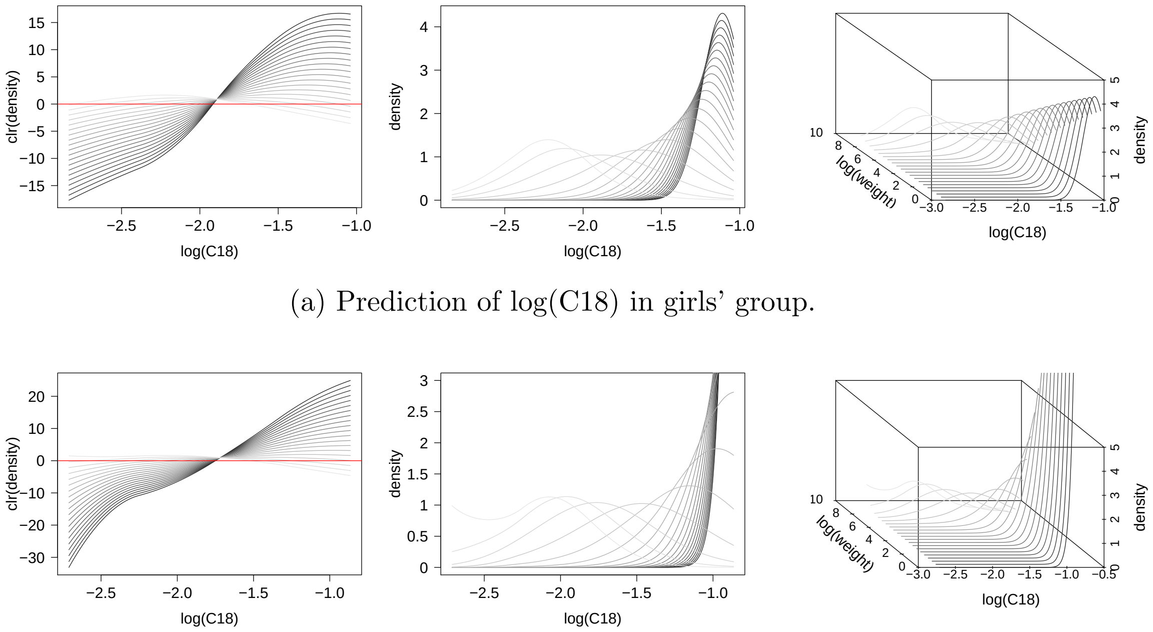
Table 3: Estimates of regression parameter vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ with marking – g for girls, b for boys (colourless rows), together with the corresponding estimates of the standard deviations $\widehat{\sigma} = \sqrt{\left\{\widehat{\text{var}}(\text{vec}(\widehat{\mathbf{B}}))\right\}_{k,k}}$ (grey rows).



(a) Estimates of regression parameters in girls' group.

(b) Estimates of regression parameters in boys' group.

Figure 18: Estimates of regression parameters. In both lines, from left to right: estimates in $L_0^2(\lambda)$ (clr transformed), estimate of β_0 in $\mathcal{B}^2(\lambda)$, estimate of β_1 in $\mathcal{B}^2(\lambda)$.



(a) Prediction of log(C18) in girls' group.

(b) Prediction of log(C18) in boys' group.

Figure 19: 2D and 3D graphs of predicted distributions of log(C18) for increasing sequence of 20 values of log weights.

vice versa. This can be better appreciated from Figure 19, 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 the 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 on the distribution of the metabolite seems to be even more dramatic. This indicates a more serious impact of the underweight to a predisposition of the metabolic disease for boys.

To assess the goodness-of-fit of the model on the observed density curves, a pointwise version of the coefficient of determination $R^2(t)$, $t \in \Omega$, was computed based on the pointwise comparison between the predicted clr transformed

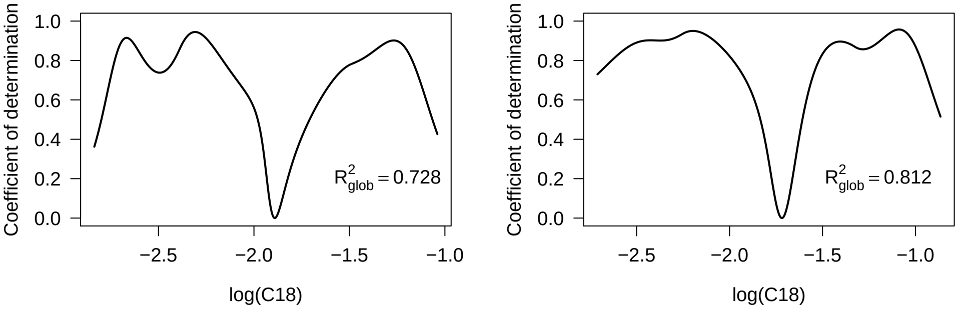


Figure 20: Pointwise coefficient of determination – on the left for girls and on the right for boys).

density and the actual data. Additionally, a global coefficient of determination, denoted by R^2_{glob}, was computed as

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

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 20. 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.

To support the interpretation of the parameters of the regression models, it is desirable to incorporate uncertainty in the 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 the re-sampling of the model-residuals. More precisely, having estimated the model, we computed the estimated residuals as $\mathrm{clr}_\lambda(\hat{\varepsilon}_i) = \mathrm{clr}_\lambda(y_i) - \mathrm{clr}_\lambda(\hat{y}_i)$. For each bootstrap repetition, we generated the bootstrap sample $\mathrm{clr}_\lambda(\varepsilon_1^{boot}), \ldots, \mathrm{clr}_\lambda(\varepsilon_N^{boot})$ by sampling with repetition from $\{\mathrm{clr}_\lambda(\hat{\varepsilon}_1), \ldots, \mathrm{clr}_\lambda(\hat{\varepsilon}_N)\}$. We defined the corresponding bootstrap response variables

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

and collect bootstrap sample for fixed predictor $\log(w_i)$ using original estimates $\beta_0(t)$ and $\beta_1(t)$, respectively, i.e.,

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

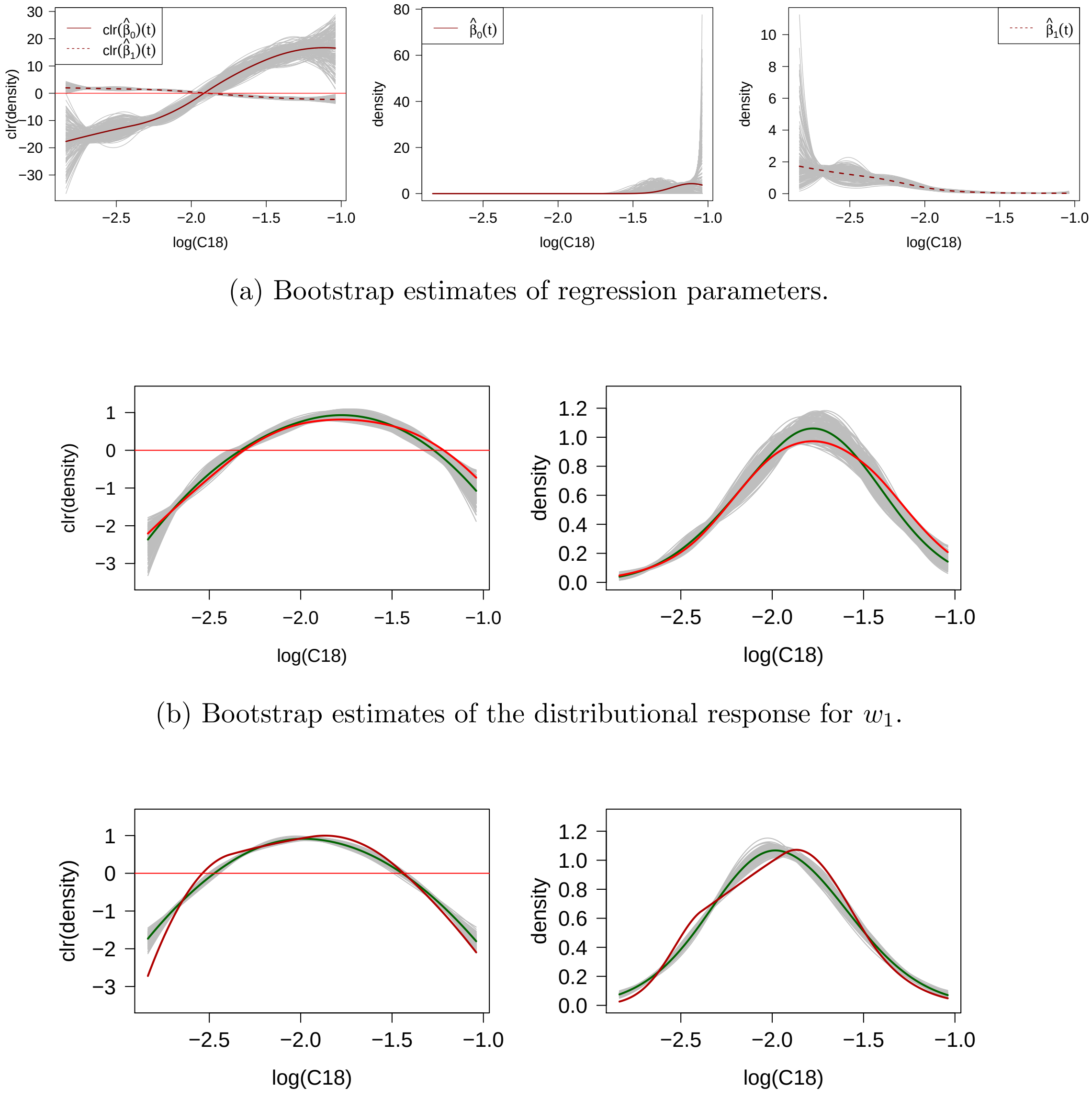
We considered $R = 200$ bootstrap repetitions, which seemed sufficient for the purpose of uncertainty assessment. For each bootstrap sample, we fitted the model and obtained the corresponding estimates of the parameters, denoted by $(\widehat{\beta}_{0r}^{boot}, \widehat{\beta}_{1r}^{boot})$, for $r = 1, ..., R$. The estimated β's and the bootstrap repetitions are displayed in Figure 21a and 22a.

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 the bootstrap procedure. The results are displayed in Figure 21 and 22, panels (b) and (c). In both cases, interesting patterns appear by observing the figures, 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. For the girls' case, Figure 21c indicates poor fitting for observed distribution corresponding to $\log(w_5)$ which can be also read from pointwise evaluated coefficient of determination (see Figure 20). This can indicate that 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

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

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

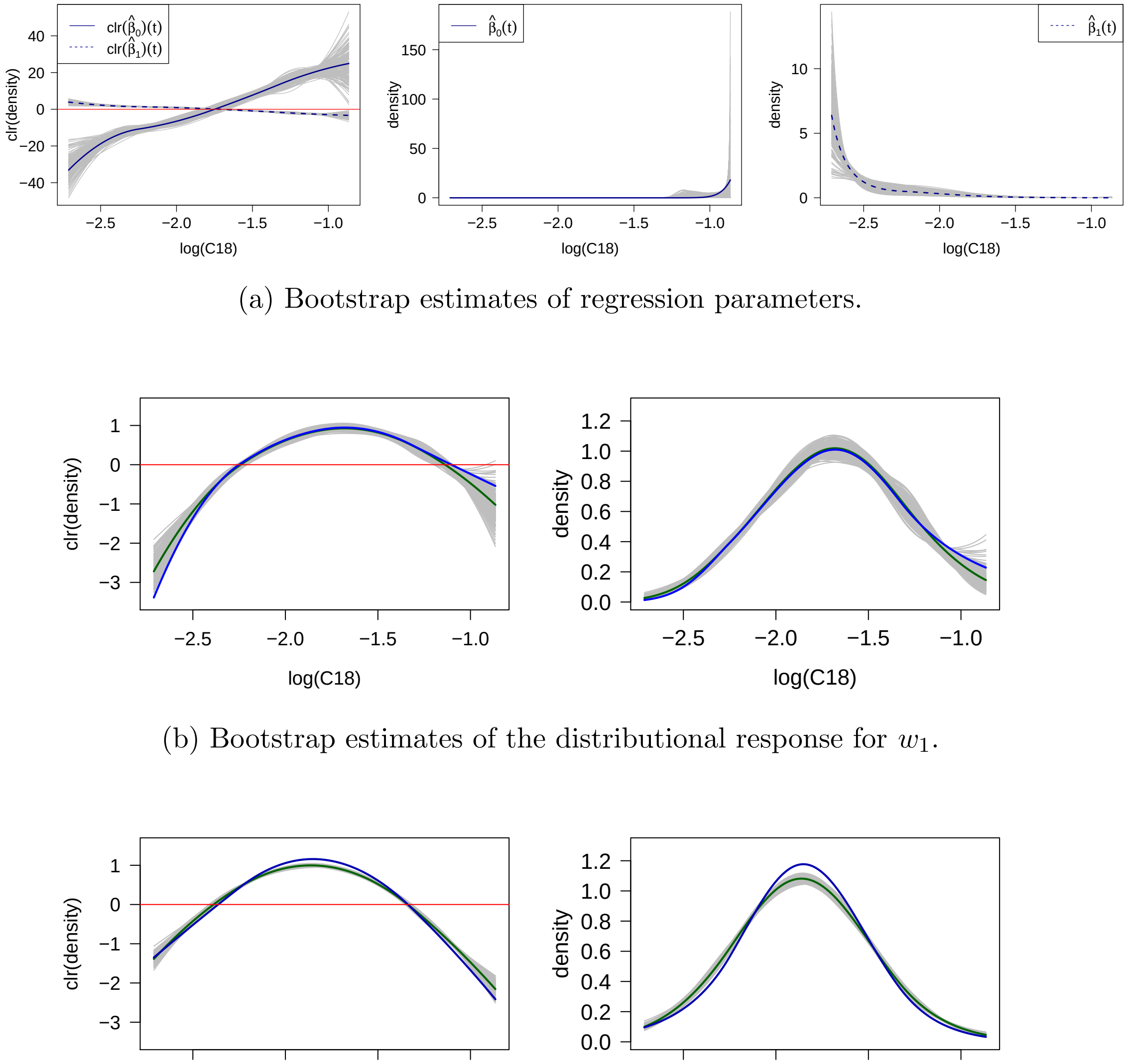


(a) Bootstrap estimates of regression parameters.

(b) Bootstrap estimates of the distributional response for w_1.

(c) Bootstrap estimates of the distributional response for w_5.

Figure 21: Bootstrap results for the girls' group. In panels (a): red curves indicate estimates of regression parameters, grey lines indicate the $R = 200$ bootstrap estimates for both the regression parameters. In panels (b) and (c): red curves indicate observed distributions for w_1 (panel (b)) and w_5 (panel (b)), green curves indicate the fitted distribution for w_1 and w_5 by model (80), grey lines indicate the corresponding fitted distributions obtained by bootstrap procedure.



(a) Bootstrap estimates of regression parameters.

(b) Bootstrap estimates of the distributional response for w_1.

(c) Bootstrap estimates of the distributional response for w_5.

Figure 22: Bootstrap results for the boys' group. In panels (a): blue curves indicate estimates of regression parameters, grey lines indicate the $R = 200$ bootstrap estimates for both the regression parameters. In panels (b) and (c): blue curves indicate observed distributions for w_1 (panel (b)) and w_5 (panel (b)), green curves indicate the fitted distribution for w_1 and w_5 by model (80), grey lines indicate the corresponding fitted distributions obtained by bootstrap procedure.

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.

4 Weighted Bayes spaces

The weighting of a domain of PDFs can be relevant in practice, as rarely all regions of the domain have the same importance or relevance for the analysis. For example, it is known that in particle-size distributions [29, 30, 31], finer fractions of soil are measured for some methods with lower reliability than crude fractions [16], which implies naturally higher relevance of the latter and their respective subdomain. Another example is represented by income distributions across various regions (see Section 5.2). The lower-income values are going to be of primary interest for policy makers when the aim is to reveal regions suffering from poverty. In addition, here the relative scale, which implies a larger impact of changes in small income values, matters and should be highlighted. And yet another reason why weighting can be convenient is to analyze deviations from a common trend in data. All these cases can benefit from a sensible weighting scheme which gives more relevance to certain regions of the domain of the PDF when conducting functional data analysis.

Weighted Bayes spaces refer to Bayes spaces with the reference measure other than the uniform one. The name *weighted Bayes spaces* reflects the fact that changing the uniform reference measure induces a (non-uniform) weighting of the domain of PDFs. Linking a weighting scheme to a non-uniform reference measure has been already discussed for multivariate compositions in [13]. As mentioned previously, rarely all regions of the domain (compositional parts in the multivariate case) have the same importance. Such weighting can be indeed relevant to consider a relative scale on the domain of a distributional variable [26]. For instance, coming back to the example of income distributions, changes in the low-income stratum (e.g. an increase of 100 € for an income of 1000 € per month) are typically of greater importance than the same absolute differences for higher earners (e.g. increase of 100 € for an income of 10,000 € per month). Accordingly, a sensible weighting strategy may be aimed at emphasizing the variability in the bottom of the domain. A weighting scheme can also be considered to account for imprecise values near the detection limit of a measurement device. The choice of the reference measure should be thus driven by the purpose of the analysis, in order to up-weight (or down-weight) subdomains (i.e., sets of parts in the discrete case) that may have greater (or lesser) importance for the analysis. More in general, data-driven weighting schemes were shown to play a crucial role for the statistical analysis in the context of domain-selection procedures, clustering, testing and regression of functional data (see [5, 18] for examples in L^2).

Following the notation from Section 1.1, we now assume that the reference measure of a measurable space $(\Omega, \mathcal{A}) = ([a,b], B([a,b]))$ is fixed to a general (probability) measure P. Then given measure μ with its P-density $f = d\mu/d\mathsf{P}$ (i.e., w.r.t. the reference measure P), the probability measure of any event $\mathrm{B} \in B([a,b])$ is

$$\mu(\mathrm{B}) = \int_{\mathrm{B}} f \, d\mathsf{P} = \int_{\mathrm{B}} \frac{d\mu}{d\mathsf{P}} \, d\mathsf{P}.$$

Note that the choice of the reference measure is not scale invariant, because it reflects on the scale of the entire Bayes space. For instance, the Lebesgue measure on a domain $\Omega = [a,b]$ is proportional to the uniform measure P_0 on Ω (hence, it belongs to the same $\mathcal{B}$-equivalence class as P_0). Clearly, λ has density $d\lambda/d\lambda = 1$ with respect to itself, whereas it has density $d\lambda/d\mathsf{P}_0 = b - a$ w.r.t. P_0. Thus, a rescaling of the reference measure determines a rescaling of the *total*. For example, when λ is considered, the total is set to $\lambda(\Omega) = b - a$, whereas P_0 is associated with a total equal to $\mathsf{P}_0(\Omega) = 1$. On the other hand, once the scale of the reference measure is fixed, the corresponding densities satisfy the scale invariance property. For instance, having set the reference measure on $\Omega = [a,b]$ to λ, the Lebesgue density $d\lambda/d\lambda$ and the uniform density $d\mathsf{P}_0/d\lambda = \frac{1}{b-a}$ are equivalent. This is further exemplified in Section 4.3 where a detailed simulation study is provided. As such, it will always be necessary to specify the total mass of the reference measure as this matters for the analysis.

Since a typical choice for P is the Lebesgue measure, restricted here to a bounded support, it opens a question on how to change the reference from λ to a measure P with strictly positive λ-density $p = d\mathsf{P}/d\lambda$. This is done by using the well-known chain rule, i.e. for a generic measure μ we have that

$$\mu(\mathrm{B}) = \int_{\mathrm{B}} \frac{d\mu}{d\lambda} \, d\lambda = \int_{\mathrm{B}} \frac{d\mu}{d\lambda} \cdot \frac{d\lambda}{d\mathsf{P}} \, d\mathsf{P} = \int_{\mathrm{B}} \frac{d\mu}{d\lambda} \cdot \frac{1}{p} \, d\mathsf{P}.$$

Given a σ-finite measure P, the Bayes space $\mathcal{B}^2(\mathsf{P})$ is a space of $\mathcal{B}$-equivalence classes of σ-finite positive measures μ with square-integrable log-density w.r.t. the reference measure P:

$$\mathcal{B}^2(\mathsf{P}) = \left\{ \mu \in \mathcal{B}^2(\mathsf{P}) : \int \left| \ln \frac{d\mu}{d\mathsf{P}} \right|^2 d\,\mathsf{P} < +\infty \right\}, \tag{81}$$

where measures are identified with the corresponding Radon-Nikodym densities; or, equivalently, $\mathcal{B}^2(\mathsf{P})$ consists of $\mathcal{B}(\mathsf{P})$-equivalence classes of proportional density functions $f = \frac{d\mu}{d\mathsf{P}}$ on $\Omega = [a, b]$ whose logarithm is square-integrable w.r.t. P.

The reason for adopting a different reference measure P can be motivated by weighting itself, but it should be also remarked that it is necessary when dealing with PDFs on possibly unbounded supports [43].

4.1 Hilbert structure of weighted Bayes spaces

In this section, we introduce a Hilbert space geometry of weighted Bayes spaces. That is, the definition of basic operations, perturbation and powering, and inner product under the general reference measure P will be considered. While both operations remain formally unchanged when changing the reference measure, the weighting affects the inner product. Here also the absolute scale of reference measure P matters, which corresponds to volume of the space Ω. It is possible to express densities from the Bayes space in the L^2 space (with respect to reference measure P) using clr transformation. This, however, still leaves open the problem of how to express the weighted densities in an *unweighted* L^2 space. A possible solution will be presented in Section 4.2.

Using a reference measure P, in [42] the operations of *perturbation* and *powering* are defined as

$$(\mu \oplus_{\mathsf{P}} \nu)(\mathrm{B}) =_{\mathcal{B}(\mathsf{P})} \int_{\mathrm{B}} \frac{d\mu}{d\mathsf{P}}(t) \cdot \frac{d\nu}{d\mathsf{P}}(t)\, d\mathsf{P}(t), \quad \mathrm{B} \in B \tag{82}$$

and

$$(\alpha \odot_{\mathsf{P}} \mu)(\mathrm{B}) =_{\mathcal{B}(\mathsf{P})} \int_{\mathrm{B}} \left(\frac{d\mu}{d\mathsf{P}}(t) \right)^{\alpha} d\mathsf{P}(t), \quad \mathrm{B} \in B, \tag{83}$$

where μ and ν are measures in $\mathcal{B}^2(\mathsf{P})$ and α is a real number. Moreover, all the measures μ, ν, λ and P are assumed to be well-defined. Consequently, these operations define a vector space structure on $\mathcal{B}^2(\mathsf{P})$ [42].

The operations (82) and (83) can be equivalently expressed using the densities with respect to P. Denoting them by $f_{\mathsf{P}} = \frac{d\mu}{d\mathsf{P}}$ and $g_{\mathsf{P}} = \frac{d\nu}{d\mathsf{P}}$ respectively, we have that

$$(f_{\mathsf{P}} \oplus_{\mathsf{P}} g_{\mathsf{P}})(t) =_{\mathcal{B}(\mathsf{P})} f_{\mathsf{P}}(t) \cdot g_{\mathsf{P}}(t) \quad \text{and} \quad (\alpha \odot_{\mathsf{P}} f_{\mathsf{P}})(t) =_{\mathcal{B}(\mathsf{P})} f_{\mathsf{P}}(t)^{\alpha}.$$

It is easy to verify that scale invariance of the reference density p holds for these operations. On the other hand, the scale of p is crucial for the definition of the inner product, defined originally in [43] and redefined here for the purpose of further developments as

$$\begin{aligned}\langle f_{\mathsf{P}}, g_{\mathsf{P}} \rangle_{\mathcal{B}(\mathsf{P})} &= \frac{1}{2\mathsf{P}(\Omega)} \int_{\Omega} \int_{\Omega} \ln \frac{f_{\mathsf{P}}(t)}{f_{\mathsf{P}}(u)} \ln \frac{g_{\mathsf{P}}(t)}{g_{\mathsf{P}}(u)} \, d\mathsf{P}(t) d\mathsf{P}(u) \\ &= \frac{1}{2\mathsf{P}(\Omega)} \int_{\Omega} \int_{\Omega} \ln \frac{f(t)}{f(u)} \ln \frac{g(t)}{g(u)} \cdot p(t) \cdot p(u) \, d\lambda(t) d\lambda(u),\end{aligned} \tag{84}$$

which endows the Bayes space $\mathcal{B}^2(\mathsf{P})$ with a separable Hilbert space structure. As a consequence, the distance between two densities $f_{\mathsf{P}}, g_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$ is obtained as

$$d_{\mathcal{B}(\mathsf{P})}(f_{\mathsf{P}}, g_{\mathsf{P}}) = \sqrt{\frac{1}{2\mathsf{P}(\Omega)} \int_{\Omega} \int_{\Omega} \left(\ln \frac{f_{\mathsf{P}}(t)}{f_{\mathsf{P}}(u)} - \ln \frac{g_{\mathsf{P}}(t)}{g_{\mathsf{P}}(u)} \right)^2 d\mathsf{P}(t) d\mathsf{P}(u).} \tag{85}$$

The reason for redefining the inner product (84) and distance (85) with respect to [43] reflects the approach presented in the multivariate case by Egozcue & Pawlowsky-Glahn [13], where the aim was to keep dominance under change of reference measure. Specifically, let p_0 be a uniform density of a measure P_0, not necessarily normalized to $\mathsf{P}_0(\Omega) = 1$, supported in an interval (or compact set) I in $\mathbb{R}$ (or $\mathbb{R}^m$), such that

$$\mathsf{P}_0(I) = \int_I p_0(t) \, dt < +\infty.$$

Let p, q be densities in $\mathcal{B}^2(\mathsf{P}_0)$ corresponding to measures P, Q such that P dominates Q, $\mathsf{P} \succ \mathsf{Q}$, that is

$$\mathsf{P}_0(t \in I : p(t) \geq q(t)) = \mathsf{P}_0(I).$$

Then, for $f_{\mathsf{P}_0}, g_{\mathsf{P}_0} \in \mathcal{B}^2(\mathsf{P}_0)$,

$$d_{\mathcal{B}(\mathsf{P})}(f_{\mathsf{P}}, g_{\mathsf{P}}) \geq d_{\mathcal{B}(\mathsf{Q})}(f_{\mathsf{Q}}, g_{\mathsf{Q}}), \tag{86}$$

where $f_{\mathsf{P}} = f_{\mathsf{P}_0} \cdot d\mathsf{P}_0/d\mathsf{P} =_{\mathcal{B}(\mathsf{P})} f_{\mathsf{P}_0} \ominus p_{\mathsf{P}_0}$ and $g_{\mathsf{P}} = g_{\mathsf{P}_0} \cdot d\mathsf{P}_0/d\mathsf{P} =_{\mathcal{B}(\mathsf{P})} g_{\mathsf{P}_0} \ominus p_{\mathsf{P}_0}$ [19]. The property (86) represents indeed the continuous counterpart to the subcompositional dominance in compositions [33]. That is, if the volume of the space $\mathsf{P}(I)$ is greater than or equal to $\mathsf{Q}(I)$ uniformly for any subinterval of I, then distances in $\mathcal{B}(\mathsf{P})$ dominate distances in $\mathcal{B}(\mathsf{Q})$. An example of this is comparing distances in a subinterval $I_1 \subseteq I$ with those in I – restrictions to subinterval corresponding to taking subcompositions [13].

Let's denote by $L_0^2(\mathsf{P})$ the closed subspace of $L^2(\mathsf{P})$ whose elements f_0 have zero integral $\int_\Omega f_0 \, d\mathsf{P} = 0$. Since the Bayes space $\mathcal{B}^2(\mathsf{P})$ is a Hilbert space, we can define an isometric isomorphism (i.e. a bijective map preserving distances) between $\mathcal{B}^2(\mathsf{P})$ and $L_0^2(\mathsf{P})$. Such a map is provided by the *centred logratio (clr)* transformation with respect to P, which is denoted by $\mathrm{clr}_{\mathsf{P}}$ and is defined for $f_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$ by [43] as

$$f_{\mathsf{P}}^c(t) = \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})(t) = \ln f_{\mathsf{P}}(t) - \frac{1}{\mathsf{P}(\Omega)} \int_\Omega \ln f_{\mathsf{P}}(u) \, d\mathsf{P}(u), \quad t \in \Omega. \tag{87}$$

Its inverse mapping to $\mathcal{B}^2(\mathsf{P})$ is obtained by using the exponential transformation, $\exp[f_{\mathsf{P}}^c](t) = \exp[\mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})](t)$, as shown in [43]. The clr representation allows to use the ordinary geometry of $L^2(\mathsf{P})$ to conduct operations of perturbation, powering, and inner product for the elements of $\mathcal{B}^2(\mathsf{P})$, while accounting for the specific features captured by the Bayes space. Indeed,

$$\mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}} \oplus_{\mathsf{P}} g_{\mathsf{P}}) = \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}}) + \mathrm{clr}_{\mathsf{P}}(g_{\mathsf{P}}), \quad \mathrm{clr}_{\mathsf{P}}(\alpha \odot f_{\mathsf{P}}) = \alpha \cdot \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})(t)$$

and

$$\langle f_{\mathsf{P}}, g_{\mathsf{P}} \rangle_{\mathcal{B}^2(\mathsf{P})} = \langle \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}}), \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}}) \rangle_{L^2(\mathsf{P})} . \tag{88}$$

In order to prove the relationship (88) (recall: $f_{\mathsf{P}}, g_{\mathsf{P}}$ are elements of $\mathcal{B}^2(\mathsf{P})$), we develop the right-hand side of (84),

$$\begin{aligned}\langle f_{\mathsf{P}}, g_{\mathsf{P}}\rangle_{\mathcal{B}(\mathsf{P})} =& \frac{1}{2\mathsf{P}(\Omega)}\int_\Omega\int_\Omega [\ln f_{\mathsf{P}}(t) - \ln f_{\mathsf{P}}(u)]\cdot[\ln g_{\mathsf{P}}(t) - \ln g_{\mathsf{P}}(u)]\ d\mathsf{P}(t)\, d\mathsf{P}(u)\\ =& \frac{1}{2\mathsf{P}(\Omega)}\left[2\int_\Omega\int_\Omega \ln f_{\mathsf{P}}(t)\cdot \ln g_{\mathsf{P}}(t)\, d\mathsf{P}(t)\, d\mathsf{P}(u)\right.\\ & \left. -2\int_\Omega\int_\Omega \ln f_{\mathsf{P}}(t)\cdot \ln g_{\mathsf{P}}(u)\, d\mathsf{P}(t)\, d\mathsf{P}(u)\right]\\ =& \int_\Omega \ln f_{\mathsf{P}}(t)\cdot \ln g_{\mathsf{P}}(t)\, d\mathsf{P}(t) - \frac{1}{\mathsf{P}(\Omega)}\int_\Omega \ln f_{\mathsf{P}}(t)\, d\mathsf{P}(t)\cdot \int_\Omega \ln g_{\mathsf{P}}(u)\, d\mathsf{P}(u),\end{aligned}$$

which truly equals the right-hand side of (88),

$$\langle \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}}), \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})\rangle_{L^2(\mathsf{P})} = \int_\Omega \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})(t)\cdot \mathrm{clr}_{\mathsf{P}}(g_{\mathsf{P}})(t)\, d\mathsf{P}(t)$$

$$\begin{aligned}=& \int_\Omega \left[\ln f_{\mathsf{P}}(t) - \frac{1}{\mathsf{P}(\Omega)}\int_\Omega \ln f_{\mathsf{P}}(u) d\mathsf{P}(u)\right]\cdot\left[\ln g_{\mathsf{P}}(t) - \frac{1}{\mathsf{P}(\Omega)}\int_\Omega \ln g_{\mathsf{P}}(u) d\mathsf{P}(u)\right] d\mathsf{P}(t)\\ =& \int_\Omega \ln f_{\mathsf{P}}(t)\cdot \ln g_{\mathsf{P}}(t)\, d\mathsf{P}(t) d\mathsf{P}(t) - \frac{2}{\mathsf{P}(\Omega)}\int_\Omega \ln f_{\mathsf{P}}(t)\, d\mathsf{P}(t)\cdot\int_\Omega \ln g_{\mathsf{P}}(u)\, d\mathsf{P}(u)\\ & + \frac{1}{\mathsf{P}^2(\Omega)}\int_\Omega\left[\int_\Omega \ln f_{\mathsf{P}}(u)\, d\mathsf{P}(u)\int_\Omega \ln g_{\mathsf{P}}(u)\, d\mathsf{P}(u)\right] d\mathsf{P}(t)\\ =& \int_\Omega \ln f_{\mathsf{P}}(t)\cdot \ln g_{\mathsf{P}}(t)\, d\mathsf{P}(t) - \frac{1}{\mathsf{P}(\Omega)}\int_\Omega \ln f_{\mathsf{P}}(t)\, d\mathsf{P}(t)\cdot \int_\Omega \ln g_{\mathsf{P}}(u)\, d\mathsf{P}(u),\end{aligned}$$

where $t, u \in \Omega$. As pointed out in the Sections 2.2 and 3, the zero integral constraint of clr transformed P-densities ($\int_\Omega \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})d\mathsf{P} = 0$) should be taken into account for any subsequent statistical analysis.

Unlike the case of [43, Sect. 4], in this work the reference measure P in $L^2_0(\mathsf{P})$ is not necessarily a probability measure, as its normalization may lead to incoherent results when restricting the analysis to a subdomain of the original domain Ω (as was shown in the discrete case [13]).

4.2 Unweighting Bayes spaces

Most methods developed for FDA rely on the assumption that functional data are embedded in the *unweighted* L^2 space. However, the clr transformation

(87) maps measures/densities in (a subspace of) a weighted space L^2 space, i.e. $L^2_0(\mathsf{P})$. Similarly, methods developed so far in Bayes spaces ground on the assumption that a uniform reference measure is considered, as for instance in Sections 2.2 and 3. A transformation mapping P-densities from $\mathcal{B}^2(\mathsf{P})$ to an unweighted counterpart of $L^2_0(\mathsf{P})$ would have the advantage of allowing the use of most FDA methods while accounting for the weighted Bayes structure of the data. Similarly, a transformation mapping P-densities from $\mathcal{B}^2(\mathsf{P})$ to an *unweighted* space $\mathcal{B}^2(\lambda)$ would allow for the use of unweighted methods to perform actual computations. In this subsection, we derive an unweighting scheme allowing one to represent the weighted Bayes space geometry in an unweighted Bayes space, as well as in an unweighted L^2 space.

We thus aim to define three mappings. Firstly, we define ω from $\mathcal{B}^2(\lambda)$ to $\mathcal{B}^2(\mathsf{P})$ as a *weighting* map associating an unweighted λ-density to a weighted P-density. Inversely, ω^{-1} is interpreted as an *unweighting* map. Similarly, we define ω_2 and its inverse ω_2^{-1} which play the same role between the unweighted and weighted L^2 spaces, i.e. $L^2(\lambda)$ and $L^2(\mathsf{P})$ respectively. Finally, we define clr_u (*unweighting clr*) such that, for $f_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$,

$$\mathrm{clr}_u(f_{\mathsf{P}} \oplus_{\mathsf{P}} g_{\mathsf{P}}) = \mathrm{clr}_u(f_{\mathsf{P}}) + \mathrm{clr}_u(g_{\mathsf{P}}), \quad \mathrm{clr}_u(\alpha \odot f_{\mathsf{P}}) = \alpha \cdot_{\mathsf{P}} t_{\mathsf{P}} \mathrm{clr}_u(f_{\mathsf{P}})(t)$$

and

$$\langle f_{\mathsf{P}}, g_{\mathsf{P}} \rangle_{\mathcal{B}^2(\mathsf{P})} = \langle \mathrm{clr}_{\mathsf{u}}(f_{\mathsf{P}}), \mathrm{clr}_{\mathsf{u}}(f_{\mathsf{P}}) \rangle_{L^2(\lambda)} . \tag{89}$$

To support this construction and study the properties of these maps, we shall use an auxiliary measure $\sqrt{\mathsf{P}}$ defined as

$$\sqrt{\mathsf{P}}(\mathrm{A}) = \int_{\mathrm{A}} \sqrt{p}\, d\lambda, \quad \mathrm{A} \in \mathcal{A}.$$

This measure plays the role of *unweighting* measure, in the sense that it allows to consistently map the *weighted* Bayes space $\mathcal{B}^2(\mathsf{P})$ into a subset of the *unweighted* L^2 space. We refer the reader to the scheme in Figure 23 as a concise representation of these relationships.

We define the $\mathcal{B}^2$-*weighting* map ω as

$$\begin{aligned} \omega : \mathcal{B}^2(\lambda) &\to \mathcal{B}^2(\mathsf{P}) \\ \varphi &\mapsto \omega(\varphi) = \varphi^{1/\sqrt{p}}, \end{aligned} \tag{90}$$

$$\begin{tikzcd}
L^2_0(\lambda) \arrow[r, shift right, "\exp"'] & \mathcal{B}^2(\lambda) \arrow[l, shift right, "\mathrm{clr}_\lambda"'] \arrow[r, shift left, "\ominus p"] & \mathcal{B}^2(\mathsf{P}) \arrow[l, shift left, "\oplus p"] \arrow[r, shift left, "\mathrm{clr}_{\mathsf{P}}"] \arrow[d, shift left, "\omega^{-1}"] \arrow[dr, "\mathrm{clr}_u"] & L^2_0(\mathsf{P}) \arrow[l, shift left, "\exp"] \arrow[d, shift left, "\omega_2^{-1}"] \\
 & & \mathcal{B}^2(\lambda) \arrow[u, shift left, "\omega"] \arrow[r, shift left, "\mathrm{clr}_{\sqrt{\mathsf{P}}}"] & L^2_{0,\sqrt{\mathsf{P}}}(\lambda) \arrow[l, shift left, "\exp"] \arrow[u, shift left, "\omega_2"]
\end{tikzcd}$$

Figure 23: Relationships among weighted and unweighted Bayes spaces, $\mathcal{B}^2(\mathsf{P})$ and $\mathcal{B}^2(\lambda)$, and weighted and unweighted $L^2(\mathsf{P})$ and $L^2(\lambda)$ spaces.

where $p = \frac{d\mathsf{P}}{d\lambda}$ (recall: p is assumed to be strictly positive in Ω). In (90), the map ω is formulated for measures, but it can be equivalently expressed using densities with respect to respective reference measures. This map defines a bijection between $\mathcal{B}^2(\lambda)$ and $\mathcal{B}^2(\mathsf{P})$, as proved in the following proposition.

Proposition 4.2.1 *The map ω defined in (90) is one-to-one and onto.*

Proof. For $\varphi \in \mathcal{B}^2(\lambda)$, let be $\omega(\varphi) = \varphi^{1/\sqrt{p}}$. Clearly, $\omega(\varphi)$ is uniquely defined. Then $\omega(\varphi) \in \mathcal{B}^2(\mathsf{P})$ due to

$$\int_\Omega \ln^2(\varphi)\, d\lambda = \int_\Omega \ln^2[\omega(\varphi)^{\sqrt{p}}]\, d\lambda = \int_\Omega \ln^2[\omega(\varphi)] p\, d\lambda = \int_\Omega \ln^2[\omega(\varphi)]\, d\mathsf{P}.$$

Inversely, for $\omega(\varphi) \in \mathcal{B}^2(\mathsf{P})$, let be $\varphi = \omega(\varphi)^{\sqrt{p}}$. Then, based on the same arguments, it results that $\varphi \in \mathcal{B}^2(\lambda)$. □

The inverse ω^{-1} is defined as $\omega^{-1}(\psi) = \psi^{\sqrt{p}}$ and it is interpreted as a $\mathcal{B}^2$-*unweighting* map. It is represented in the bottom left part of the scheme in Figure 23. Obviously, both ω and ω^{-1} depend on the scale of P.

We define the L^2-*weighting* map ω_2 as

$$\begin{aligned} \omega_2 : L^2(\lambda) &\to L^2(\mathsf{P}) \\ \eta &\mapsto \omega(\eta) = \eta/\sqrt{p}. \end{aligned}$$

Using the same rationale as for Proposition 4.2.1, it can be proved that ω_2 defines a bijection between $L^2(\lambda)$ and $L^2(\mathsf{P})$. Its inverse ω_2^{-1} is defined as $\omega_2^{-1}(\xi) = \xi\sqrt{p}$ and it is interpreted as a L^2-*unweighting* map. It is represented in the bottom right part of the scheme in Figure 23. Note that ω is non-linear with respect to the Bayes space geometry, as well as ω_2 is non-linear in L^2.

Using (84), the map $\mathrm{clr}_u : \mathcal{B}^2(\mathsf{P}) \to L^2(\lambda)$ can be then defined as

$$\mathrm{clr}_u(f_{\mathsf{P}}) = \omega_2^{-1}[\mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})]. \tag{91}$$

It can be proven that (91) fulfills all the properties detailed in (89). Note that the scale of clr_u depends on the scale of $\sqrt{p}$, hence on the scale of $\sqrt{\mathsf{P}}$, because of the non-linearity of ω_2 (see [10] for the case of finite-dimensional compositions). As such, similarly to the multivariate case [13], the scale of the reference measure is relevant in the geometry of both weighted and unweighted spaces.

It is worth noticing that clr_u is closely related to a different centered logratio transformation. This is defined on the unweighted space $\mathcal{B}^2(\lambda)$ and induced by the unweighting measure $\sqrt{\mathsf{P}}$. Indeed, let $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ be the subspace of $L^2(\lambda)$ such that $\int_\Omega f \, d\sqrt{\mathsf{P}} = 0$ for $f \in L^2(\lambda)$. Let's define on $\mathcal{B}^2(\lambda)$ the map $\mathrm{clr}_{\sqrt{\mathsf{P}}}$ as

$$\mathrm{clr}_{\sqrt{\mathsf{P}}}(\varphi)(t) = \ln \varphi(t) - \frac{1}{\sqrt{\mathsf{P}}(\Omega)} \int_\Omega \ln[\varphi(u)] \, d\sqrt{\mathsf{P}}(u), \quad t \in \Omega, \quad \varphi \in \mathcal{B}^2(\lambda). \tag{92}$$

In light of Proposition 4.2.1, it is easy to see that the map (92) is well defined. For any $\varphi \in \mathcal{B}^2(\lambda)$, we can set $f_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$ to $f_{\mathsf{P}} = \omega(\varphi) = \varphi^{1/\sqrt{p}}$. Then, it holds that

$$\int_\Omega \ln[\varphi(u)] \, d\sqrt{\mathsf{P}}(u) = \int_\Omega \ln[f_{\mathsf{P}}(u)] p(u) \, d\lambda(u) < +\infty.$$

Moreover, for any φ in $\mathcal{B}^2(\lambda)$, we have that $\mathrm{clr}_{\sqrt{\mathsf{P}}}(\varphi) \in L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$. The following proposition establishes the close relationship between clr_u and $\mathrm{clr}_{\sqrt{\mathsf{P}}}$, thus completing the scheme in Figure 23.

Proposition 4.2.2 *The following statements hold true.*

(i) The image of the space $\mathcal{B}^2(\mathsf{P})$ under the map clr_u defined in (91) is $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$.

(ii) The map clr_u *coincides with the composed function* $\mathrm{clr}_{\sqrt{\mathsf{P}}} \circ \omega^{-1}$*, i.e.*

$$\mathrm{clr}_u(f_{\mathsf{P}}) = \mathrm{clr}_{\sqrt{\mathsf{P}}}(\omega^{-1}(f_{\mathsf{P}})) \quad \textit{and} \quad f_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P}).$$

(iii) The inverse of the map $\mathrm{clr}_{\sqrt{\mathsf{P}}}$ *is* $\mathrm{clr}^{-1}_{\sqrt{\mathsf{P}}} : L^2_{0,\sqrt{\mathsf{P}}}(\lambda) \to \mathcal{B}^2(\lambda)$ *and is given by*

$$\mathrm{clr}^{-1}_{\sqrt{\mathsf{P}}}(\psi) =_{\mathcal{B}^2(\lambda)} \exp(\psi),$$

for any ψ *in* $L^2_{0,\sqrt{\mathsf{P}}}$.

(iv) The inverse of the map clr_u *is* $\mathrm{clr}^{-1}_u : L^2_{0,\sqrt{\mathsf{P}}}(\lambda) \to \mathcal{B}^2(\mathsf{P})$ *and is given by*

$$\mathrm{clr}^{-1}_u(\psi) =_{\mathcal{B}^2(\mathsf{P})} \exp[\omega_2(\psi)] =_{\mathcal{B}^2(\mathsf{P})} \omega[\exp(\psi)],$$

for any ψ *in* $L^2_{0,\sqrt{\mathsf{P}}}$.

Proof. Statement (i). Let's denote by f_{P} a density in $\mathcal{B}^2(\mathsf{P})$. Then

$$\int_\Omega \mathrm{clr}_u(f_{\mathsf{P}})\, d\sqrt{\mathsf{P}} = \int_\Omega \mathrm{clr}_u(f_{\mathsf{P}})\, \sqrt{p}\, d\lambda = \int_\Omega \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})\, d\mathsf{P} = 0,$$

proving the first statement.

Statement (ii). Consider $f_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$. Then

$$\begin{aligned}\mathrm{clr}_{\sqrt{\mathsf{P}}}(\omega^{-1}(f_{\mathsf{P}})) &= \ln(f_{\mathsf{P}}^{\sqrt{p}}) - \frac{1}{\sqrt{\mathsf{P}}(\Omega)} \int_\Omega \ln(f_{\mathsf{P}}^{\sqrt{p}})\, d\sqrt{\mathsf{P}} = \\ &= \sqrt{p}\ln(f_{\mathsf{P}}) - \frac{1}{\sqrt{\mathsf{P}}(\Omega)} \int_\Omega \ln(f_{\mathsf{P}}) p\, d\lambda. \qquad (93)\end{aligned}$$

Let's call $\xi \in L^2_0(\mathsf{P})$ the element $\xi = \mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})$. Since $\mathrm{clr}_{\mathsf{P}}$ is one-to-one and onto between $\mathcal{B}^2(\mathsf{P})$ and $L^2_0(\mathsf{P})$, it holds that $f_{\mathsf{P}} =_{\mathcal{B}^2(\mathsf{P})} \exp \xi$ and we can rewrite

$$\mathrm{clr}_{\sqrt{\mathsf{P}}}(\omega^{-1}(f_{\mathsf{P}})) = \xi\sqrt{p},$$

where the last term of (93) cancels because $\xi \in L^2_0(\mathsf{P})$. Considering clr_u, using the same notation as before, it results that

$$\mathrm{clr}_u(f_{\mathsf{P}}) = \sqrt{p} \cdot \left[\ln(f_{\mathsf{P}}) - \frac{1}{\mathsf{P}(\Omega)} \int_\Omega \ln(f_{\mathsf{P}})\, d\mathsf{P}\right] = \xi\sqrt{p}.$$

Statement (iii). For $\psi \in L^2_{0,\sqrt{\mathsf{P}}}$, it holds that

$$\mathrm{clr}_{\sqrt{\mathsf{P}}}[\exp(\psi)](u) = \ln[\exp(\psi)] - \frac{1}{\sqrt{\mathsf{P}}(\Omega)} \int_{\Omega} \ln[\exp(\psi(u))] d\sqrt{\mathsf{P}}(u) = \psi(u),$$

for any $u \in \Omega$.

Statement (iv). This is an obvious consequence of the previous point (iii). □

Note that taking the $\mathcal{B}^2$-*unweighting* transformation ω^{-1} is indeed different from simply changing the reference measure from P to λ. The former transformation is indeed used to *represent* the weighted Bayes space through an unweighted one, while preserving its weighted Hilbert geometry. In fact, as further highlighted in Section 5.2, this auxiliary space may serve to enhance the interpretation of the weighted structure. For instance, visual interpretation of a weighted density f_{P} in $\mathcal{B}^2(\mathsf{P})$ is hindered by the need to take into account the weighting scheme considered for the support. On the contrary, visualisation of the corresponding unweighted density $\omega^{-1}(f_{\mathsf{P}})$ allows for the usual interpretation, yet representing the same object – just by incorporating the weighting scheme.

As a way of example, consider the graphs in Figure 24, whose layout recalls that of Figure 23. Figure 24b represents a uniform density with respect to the Lebesgue measure (its density $d\lambda/d\lambda = 1$ is displayed as a grey line), on the interval $\Omega = [0, 0.5]$, i.e., $f(t) = 2$, $t \in \Omega$. Such density is embedded in $\mathcal{B}^2(\lambda)$; an equivalent representation is its clr transformation $\mathrm{clr}_{\lambda}(f)$, which is an element of $L^2_0(\lambda)$ (in fact, $\mathrm{clr}(f)(t) = 0$, $t \in \Omega$) and is reported in Figure 24a. To give higher relevance to the right part of the domain with respect to the left one, one may consider as a weighting scheme that induced by the function $p(t)$ displayed as a grey line in Figure 24c. Here, $p(t) = d\mathsf{P}/d\lambda$ is defined as a Gaussian density with the mean $\mu = 0.5$ and the standard deviation $\sigma = 0.2$, truncated to the interval Ω, and normalized over this interval. The density f, when changing the reference measure to P (i.e., $f_{\mathsf{P}} = f \ominus p$), is no longer uniform; it is displayed as a black curve in Figure 24c. Note that, when interpreting Figure 24c, (e.g., to compare the mass attributed to subdomains), one should pay close attention to the weighting scheme. The (weighted) density f_{P} can be represented as an element of $L^2(\mathsf{P})$ (black curve Figure 24d) if transformed via $\mathrm{clr}_{\mathsf{P}}$. Even though such representation

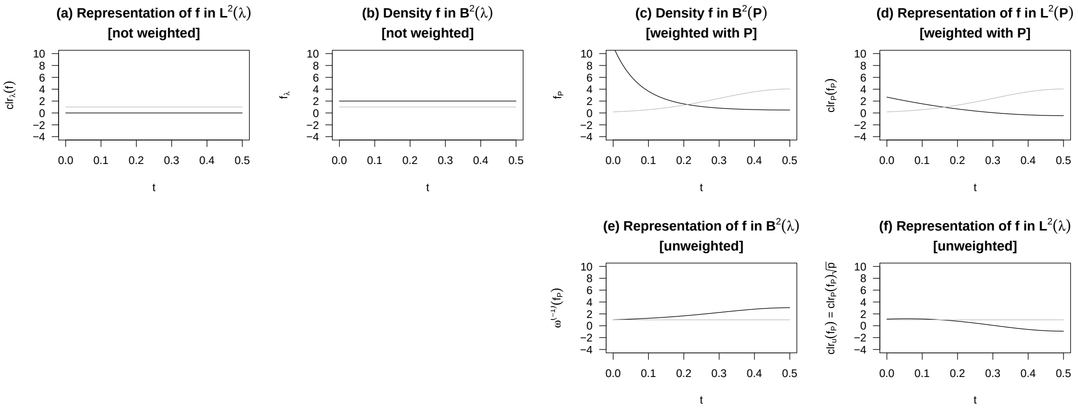


Figure 24: Changing the reference measure to the uniform density.

allows mapping the *relative* scale of $\mathcal{B}^2(\mathsf{P})$ to the *absolute* scale of $L^2(\mathsf{P})$, it still preserves the weighting scheme: in Figure 24d, the geometry still needs to account for the weights $p(t)$ (grey line). Figure 24e and 24f display the unweighted counterparts of Figure 24c and 24d, i.e., $\omega^{-1}(f_{\mathsf{P}})$ and $\omega_2^{-1}(\mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})) = \mathrm{clr}_u(f_{\mathsf{P}})$, respectively. Note that they still represent the same weighted density f_{P} but in a different (unweighted) space, which can be used for the purpose of analysis and interpretation. For instance, Figure 24e shows that the density f_{P} can be interpreted as a measure giving relatively more mass to the right part of the domain than to its left part. Such interpretation would be otherwise hard to argue from Figure 24c. Finally, note that the densities f (black line in Figure 24b) and $\omega^{-1}(f_{\mathsf{P}})$ (black line in Figure 24e) are markedly different, being in fact representatives of different weighting schemes.

It is also clear that, as long as the Lebesgue reference measure is concerned ($\mathsf{P}(\Omega) = \lambda([a,b])$), the transformations clr_u and $\mathrm{clr}_{\mathsf{P}}$ coincide, and they reduce to the clr transformation clr_λ (13). Note, however, that this would not be true for reference measures proportional to the Lebesgue one, because the scale of the reference does have an impact on the Hilbert geometry.

The above considerations have a direct impact on applications. For a sample of densities $f_1, \ldots, f_N$ to be analyzed with respect to a reference measure P, the following strategy can be adopted:

1. Set the reference measure P.

2. If the PDFs were given w.r.t. the Lebesgue measure, change the reference measure from λ to P. That is, set $f_{\mathsf{P},i} = f_i \ominus p$, for $i = 1, \ldots, N$, with $f_{\mathsf{P},i} \in \mathcal{B}^2(\mathsf{P})$.

3. Map $f_{\mathsf{P},i}$, for $i = 1, \ldots, N$, onto $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ by using the clr_u transformation. Set $y_i = \mathrm{clr}_u(f_{\mathsf{P},i})$, for $i = 1, \ldots, N$.

4. Perform the statistical analysis on y_i, $i = 1, \ldots, N$, using *unweighted* L^2_0 $(L^2_{0,\sqrt{\mathsf{P}}}(\lambda))$ methods.

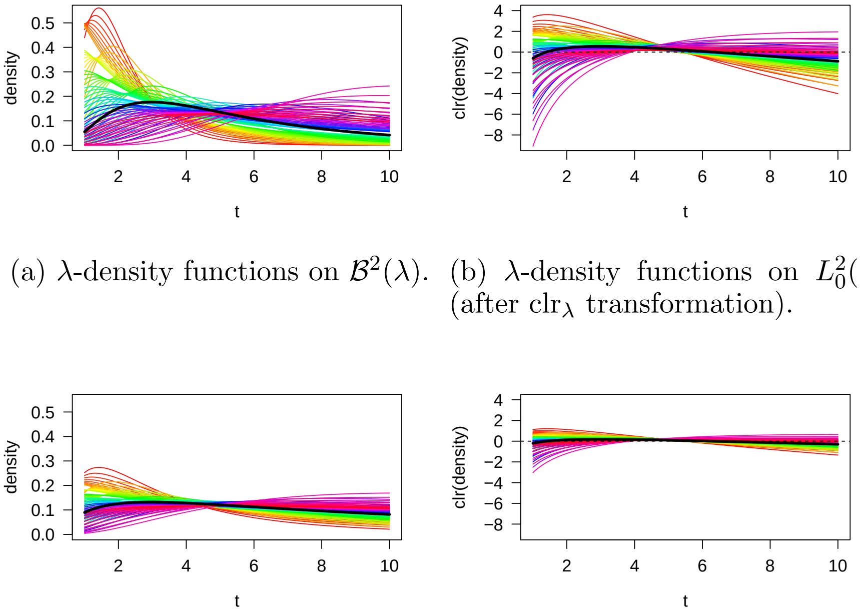
5. If the results needs to be given in terms of densities, use the inverse transformation $\exp[\mathrm{clr}_u(f_{\mathsf{P}})]$ to express the results in the unweighted space $\mathcal{B}^2(\lambda)$, where they can be easily interpreted.

This strategy is further illustrated in Section 5, which presents a dimensionality reduction method in weighted Bayes spaces.

4.3 Changing the reference measure: the consequences for density data using simulated densities from exponential families

To examine the effects of changing the reference measure, we simulate densities from two exponential families and analyze them with respect to different reference measures – Lebesgue, uniform (as its normalized counterpart) and exponential measures. While the first two reference measures represent equal weighting on the respective domains, the last one is an example of down-weighting the right-hand side of domain, possibly to stress the relative scale along the domain of the data as motivated by the income data application mentioned at the beginning of the chapter.

Inspired by the case study presented in Section 5.2, we consider a set of (truncated) log-normal densities with means $\mu_i = 0.6 + 0.25 \cdot (i-1)$ and standard deviations $\sigma_j = 0.5 + 0.07 \cdot (j-1)$ for $i, j = 1, \ldots, 9$, on the interval $\Omega = [1, 10]$. They are represented with respect to the Lebesgue measure and displayed in Figure 25a, where the color scale follows the index $\kappa = j + 9(i-1)$, $i, j = 1, ..., 9$ (i.e. equal mean values are represented with similar colors). In this case, the



(a) λ-density functions on $\mathcal{B}^2(\lambda)$.

(b) λ-density functions on $L_0^2(\lambda)$ (after clr_λ transformation).

(c) $\mathcal{B}^2$-unweighted P_0-density functions on $\mathcal{B}^2(\lambda)$.

(d) P_0-density functions on $L^2_{0,\sqrt{\mathsf{P}_0}}(\lambda)$ (after clr_u transformation).

Figure 25: Log-normal density functions w.r.t. the Lebesgue measure (panels (a)-(b)) and w.r.t. the uniform measure P_0 (panels (c)-(d)), with parameters $\mu_i = 0.6 + 0.25 \cdot (i-1)$ and $\sigma_j = 0.5 + 0.07 \cdot (j-1)$ for $i, j = 1, \ldots, 9$, $\Omega = [1, 10]$. Black curves indicate the corresponding mean functions.

transformations $\mathrm{clr}_{\mathsf{P}}$ and clr_u coincide (Figure 25b), and they are obtained as

$$\begin{aligned} f^c_\lambda(t; \mu_i, \sigma_j) = &-\frac{\ln^2 t}{2\sigma_j^2} + \left(-1 + \frac{\mu_i}{\sigma_j^2}\right)\left(\ln t - \frac{10}{9} \cdot \ln 10 + 9\right) + \\ &+ \frac{1}{\sigma_j^2}\left(1 + \frac{5}{9} \cdot \ln^2 10 - \frac{9}{10} \ln 10\right),\ t \in \Omega. \end{aligned} \tag{94}$$

To appreciate the influence of changing the scale of the reference measure, we set P_0 to be the uniform measure on Ω, $\mathsf{P}_0 = \lambda/9$ (with density $p_0(t) = 1/9$, for $t \in \Omega$). The log-normal densities w.r.t. P_0 are proportional to those in Figure 25a, which is precisely the scaling effect induced by the reference measure. The $\mathrm{clr}_{\mathsf{P}_0}$ representations of the P_0-densities coincide with those in Figure 25b; however, the former are embedded in $L^2(\mathsf{P}_0)$, whereas the latter do so in $L^2(\lambda)$. As such, a different scale is actually characterizing the two Bayes spaces. The

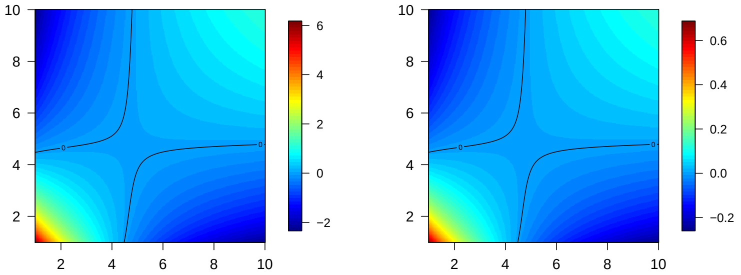


Figure 26: Covariance functions of log-normal λ-densities (left) and log-normal P_0-densities (right). To appreciate the similarity between covariance structures, colors are *not* given on the same scale.

clr_u transformed densities, i.e. $y_i = (1/3) \cdot \mathrm{clr}_{\mathsf{P}_0}(f_{\mathsf{P}_0,i})$ – which is an element of $L^2_{0,\sqrt{\mathsf{P}_0}}$ – are displayed in Figure 25d. Here, the different scales of the two spaces are apparent. Finally, Figure 25c displays the $\mathcal{B}^2$-unweighted densities, i.e., $\omega^{-1}(f_{\mathsf{P}_0,i}) = (f_{\mathsf{P}_0,i})^3$, which are now elements of $\mathcal{B}^2(\lambda)$. A graphical representation like in Figure 25c may be very convenient in applications, as it allows to visually neglect the weighting of the domain when observing the figure.

Visual inspection of Figure 25c suggests that the scaling of the reference measure by $\alpha > 1$ (or $\alpha < 1$) results in a shrinkage (or expansion) of the corresponding Bayes space. The shrinkage of the Bayes space can be readily observed by comparing Figures 25b and 25d (note that these representations are comparable because they are referred to the same reference λ). This is also well reflected in the covariance functions (Figure 26); indeed, the covariance structure is preserved but it differs in the scale. Here, the variability of the data, when these are embedded in $\mathcal{B}^2(\lambda)$ (resp. $\mathcal{B}^2(\mathsf{P}_0)$), is concentrated on the boundaries of the domain Ω. Particularly being more dominant in its left-hand side, where the densities display larger relative differences. Analogous conclusion can be derived from Figures 25b and 25d respectively, but note that these graphs are interpreted in terms of absolute differences among curves in agreement with the L^2 geometry considered therein.

For the same log-normal densities, an exponential reference measure P^δ was also considered, setting their densities to $p^\delta(t) =_{\mathcal{B}(\lambda)} \exp\{-\delta \cdot t\}, t \in \Omega$, with δ in $\{0.25, 0.75, 1.25\}$. Note that, for increasing values of δ, the reference gives increasing weight to the left-hand side of the domain Ω. In order to obtain comparable results in terms of scales, the reference measures were all considered as normalized to unity. Figure 27 depicts the resulting log-normal densities w.r.t. P,

$$f_{\mathsf{P}}(t;\mu_i,\sigma_j) =_{\mathcal{B}(\mathsf{P})} \frac{1}{t}\cdot \exp\left\{-\frac{\ln t-\mu_i}{2\sigma_j^2}+\delta\cdot t\right\}, \quad t\in\Omega,$$

as well as their counterparts in $L_0^2(\mathsf{P})$ and $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$. As expected, by downweighting the right-hand side of the domain (i.e. increasing δ), the variability in the tails on the right is eventually completely masked, whereas the opposite trend can be observed in the tails on the left. This is apparent when comparing the log-normal densities (Figure 27e) and the corresponding covariance functions (Figures 26 and 28).

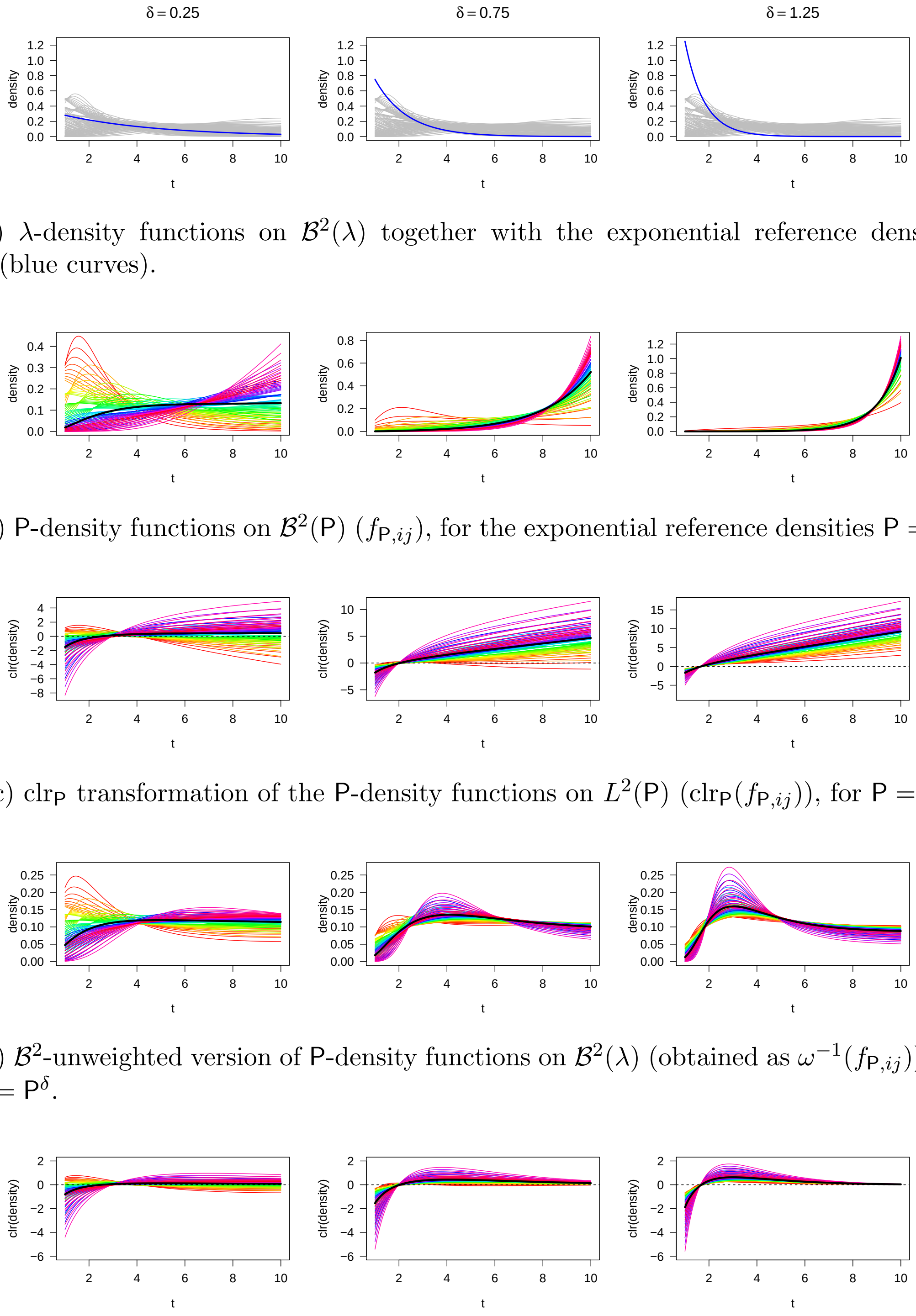
To see the weighting effects on densities whose major source of (relative) variability is in the right-hand side of domain, truncated Weibull densities with the support on $\Omega=[1,8]$ are considered, namely

$$f_{\lambda}(t;\nu_i,\theta_j) =_{\mathcal{B}(\lambda)} \left(\frac{t}{\nu_i}\right)^{\theta_j-1}\cdot \exp\left\{-\left(\frac{t}{\nu_i}\right)^{\theta_j}\right\}, \qquad t\in\Omega,$$

with shape and scale parameters set to $\theta_j = 1+0.1\cdot(j-1)$ and $\nu_i = 1+0.6\cdot(i-1)), i,j=1,\ldots,10$ respectively. The main source of variability in the latter densities is primarily displayed in the right part of the boundary, as shown in Figure 30a and Figure 30b. Figure 29 illustrates the behavior of densities when the reference measure is changed to an exponential measure $\mathsf{P}=\mathsf{P}^{\delta}$ with $\delta\in\{0.25,0.75,1.50\}$,

$$f_{\mathsf{P}}(t;\nu_i,\theta_j) =_{\mathcal{B}(\mathsf{P})} \frac{f_{\lambda}(t;\nu_i,\theta_j)}{p^{\delta}(t)} =_{\mathcal{B}(\mathsf{P})} \left(\frac{t}{\nu_i}\right)^{\theta_j-1}\cdot \exp\left\{-\left(\frac{t}{\nu_i}\right)^{\theta_j}+\delta t\right\}, \qquad t\in\Omega. \tag{95}$$

As before, the variability driven by the right tails of the distribution is reduced and becomes significantly higher in the left. In addition, for an extreme weighting ($\delta=1.50$) some differences are evidenced even in the central part of distributions (see also Figure 30, where the covariance structure of weighted densities is reported).



(a) λ-density functions on $\mathcal{B}^2(\lambda)$ together with the exponential reference densities P^δ(blue curves).

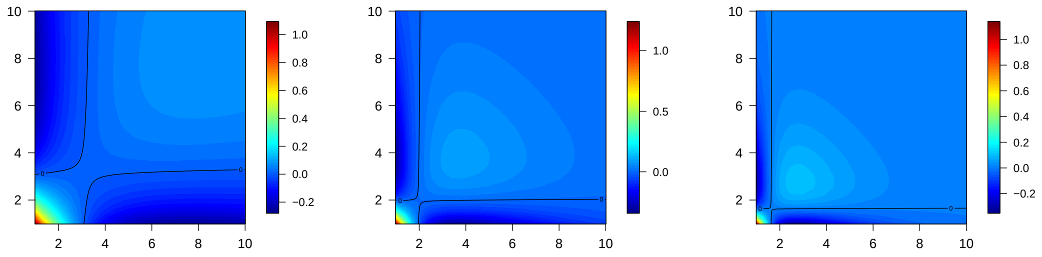
(b) P-density functions on $\mathcal{B}^2(\mathsf{P})$ $(f_{\mathsf{P},ij})$, for the exponential reference densities $\mathsf{P} = \mathsf{P}^\delta$.

(c) $\mathrm{clr}_{\mathsf{P}}$ transformation of the P-density functions on $L^2(\mathsf{P})$ $(\mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P},ij}))$, for $\mathsf{P} = \mathsf{P}^\delta$.

(d) $\mathcal{B}^2$-unweighted version of P-density functions on $\mathcal{B}^2(\lambda)$ (obtained as $\omega^{-1}(f_{\mathsf{P},ij})$), for $\mathsf{P} = \mathsf{P}^\delta$.

(e) clr_u transformation of P-density functions in $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ (obtained as $\mathrm{clr}_u(f_{\mathsf{P},ij})$), for $\mathsf{P} = \mathsf{P}^\delta$.

Figure 27: Log-normal density functions with respect to exponential reference measures with $\delta = 0.25$ (first column), $\delta = 0.75$ (second column) and $\delta = 1.25$ (third column) for parameters $\mu_i = 0.6 + 0.25 \cdot (i-1)$ and $\sigma_j = 0.5 + 0.07 \cdot (j-1)$ for $i, j = 1, \ldots, 9$ on $\Omega = [1, 10]$.

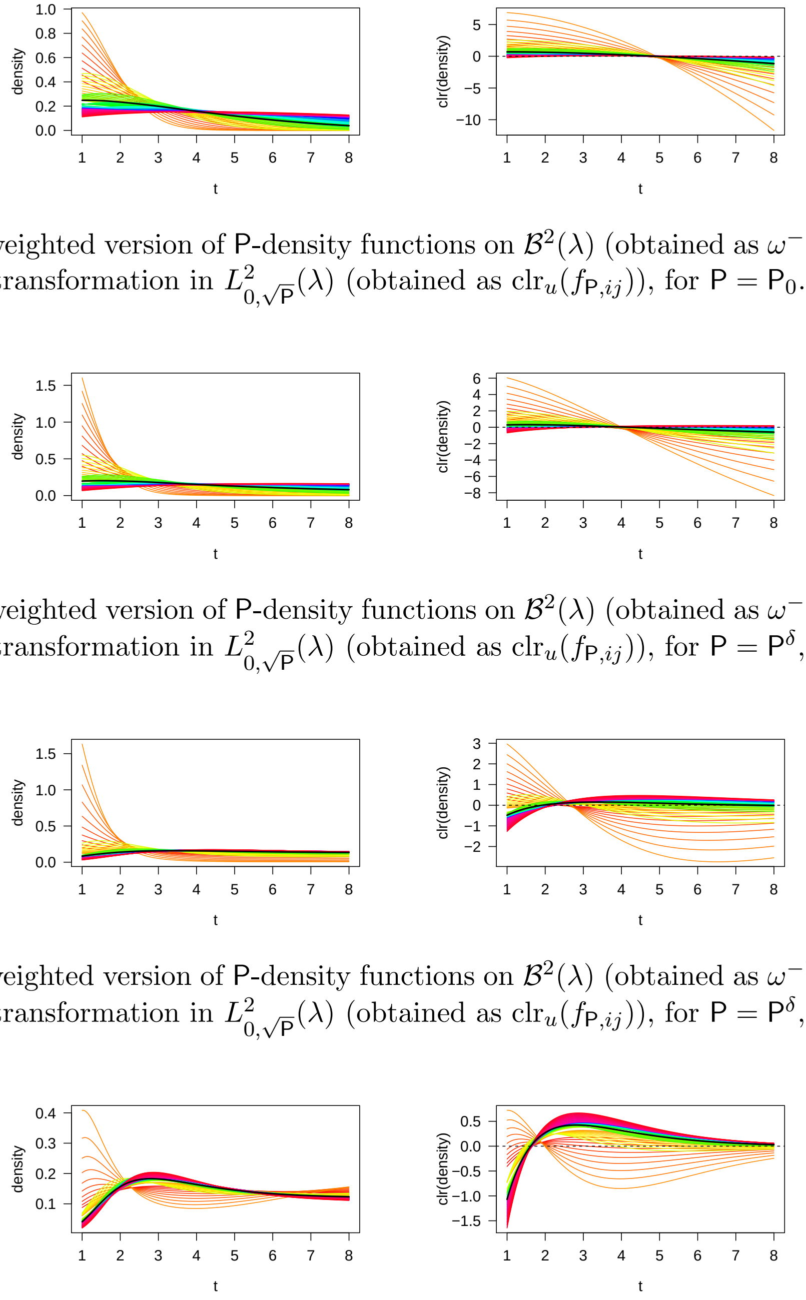


(a) P^{δ}, $\delta = 0.25$

(b) P^{δ}, $\delta = 0.75$

(c) P^{δ}, $\delta = 1.25$

Figure 28: Comparison of covariance functions for log-normal densities w.r.t. the exponential reference measure for different values of parameter δ. To appreciate the patterns of the covariance structures, colors are *not* given on the same scale.



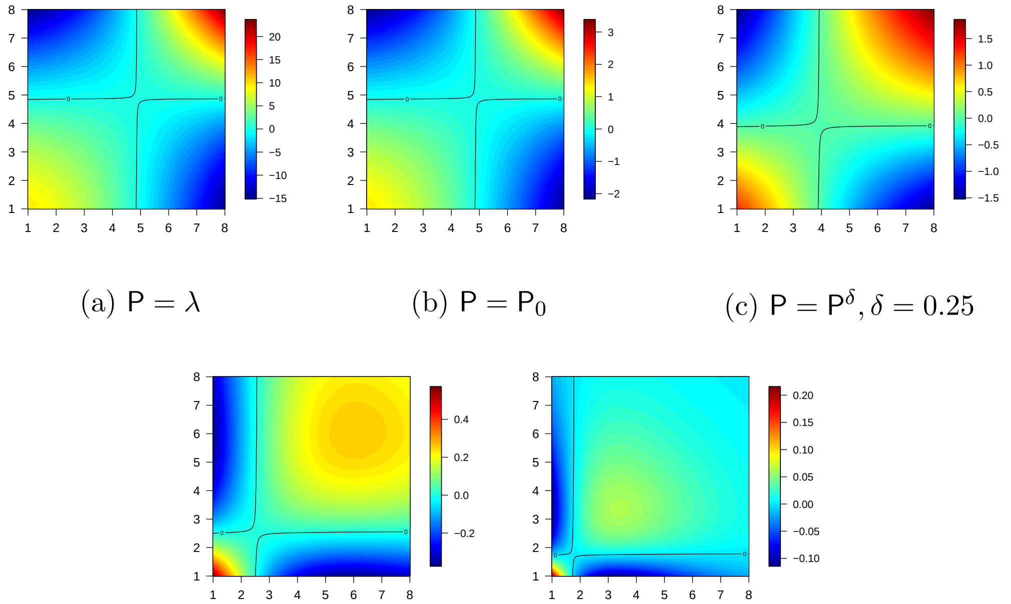
(a) $\mathcal{B}^2$-unweighted version of P-density functions on $\mathcal{B}^2(\lambda)$ (obtained as $\omega^{-1}(f_{\mathsf{P},ij})$) and their clr_u-transformation in $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ (obtained as $\mathrm{clr}_u(f_{\mathsf{P},ij})$), for $\mathsf{P} = \mathsf{P}_0$.

(b) $\mathcal{B}^2$-unweighted version of P-density functions on $\mathcal{B}^2(\lambda)$ (obtained as $\omega^{-1}(f_{\mathsf{P},ij})$) and their clr_u-transformation in $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ (obtained as $\mathrm{clr}_u(f_{\mathsf{P},ij})$), for $\mathsf{P} = \mathsf{P}^\delta$, $\delta = 0.25$.

(c) $\mathcal{B}^2$-unweighted version of P-density functions on $\mathcal{B}^2(\lambda)$ (obtained as $\omega^{-1}(f_{\mathsf{P},ij})$) and their clr_u-transformation in $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ (obtained as $\mathrm{clr}_u(f_{\mathsf{P},ij})$), for $\mathsf{P} = \mathsf{P}^\delta$, $\delta = 0.75$.

(d) $\mathcal{B}^2$-unweighted version of P-density functions on $\mathcal{B}^2(\lambda)$ (obtained as $\omega^{-1}(f_{\mathsf{P},ij})$) and their clr_u-transformation in $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ (obtained as $\mathrm{clr}_u(f_{\mathsf{P},ij})$), for $\mathsf{P} = \mathsf{P}^\delta$, $\delta = 1.50$.

Figure 29: Weibull density functions in case of (1) uniform measure ($\mathsf{P} = \mathsf{P}_0$) and (2) exponential reference measures ($\mathsf{P} = \mathsf{P}^\delta, \delta \in \{0.25, 0.75, 1.50\}$) for parameters $\theta_j = 1 + 0.1 \cdot (j - 1)$ and $\nu_i = 1 + 0.6 \cdot (i - 1)$ for $i, j = 1, \ldots, 10$ on $\Omega = [1, 8]$. The black curves indicate respective mean functions.



(a) $\mathsf{P} = \lambda$

(b) $\mathsf{P} = \mathsf{P}_0$

(c) $\mathsf{P} = \mathsf{P}^{\delta}, \delta = 0.25$

(d) $\mathsf{P} = \mathsf{P}^{\delta}, \delta = 0.75$

(e) $\mathsf{P} = \mathsf{P}^{\delta}, \delta = 1.50$

Figure 30: Comparison of covariance functions for simulated Weibull densities w.r.t. Lebesgue, uniform and the exponential reference measure for different values of parameter δ.

5 Statistical methods in weighted Bayes spaces: weighted SFPCA

Simplicial functional principal component analysis (SFPCA, [21]) was recently introduced to adapt the well-known functional principal component analysis [36] to density functions. It is grounded on the theory of Bayes spaces and assumes that the Lebesgue measure is set as a reference measure. SFPCA aims to explore the main modes of *relative* variability in a sample of density data and can be used to suggest a possible dimensionality reduction of a dataset of PDFs. In this section, we extend the SFPCA to its weighted version, named hereafter wSFPCA. Besides its relevance in applications, this extension serves as an illustrative example of the strategy detailed in Section 4.2.

Let's denote by $f_1, \dots, f_N$ an i.i.d. sample in $\mathcal{B}^2(\lambda)$. After selecting the reference measure P with λ-density p, a sample $f_{\mathsf{P},i} = f_i \ominus p$, for $i = 1, \dots, N$, in $\mathcal{B}^2(\mathsf{P})$ is obtained. We assume without loss of generality this sample is mean-centered. If this is not the case, it is enough to consider $\tilde{f}_{\mathsf{P},i} = f_{\mathsf{P},i} \ominus \bar{f}_{\mathsf{P}}$, where $\bar{f}_{\mathsf{P}}$ stands for the (weighted) sample mean of the observed (weighted) densities

$$\bar{f}_{\mathsf{P}} = \frac{1}{N} \odot_{\mathsf{P}} \bigoplus\nolimits_{\mathsf{P}_{i=1}}^{N} f_{\mathsf{P},i}.$$

Note that the centering operation shifts the center of the sample to the neutral element of the (weighted) perturbation operation, that is, the uniform density on $\mathcal{B}^2(\mathsf{P})$.

The aim of wSFPCA is to identify a collection of orthogonal and normalized P-density functions $\{\xi_{\mathsf{P},j}\}_{j\geq 1}$ in $\mathcal{B}^2(\mathsf{P})$ corresponding to the directions in $\mathcal{B}^2(\mathsf{P})$ along which the dataset displays its main modes of variability. These directions are called weighted simplicial functional principal components (wSFPCs), and they are obtained by maximizing the following objective function

$$\sum_{i=1}^{N} \langle f_{\mathsf{P},i}, \xi_{\mathsf{P}} \rangle^2_{\mathcal{B}(\mathsf{P})} \text{ subject to } \|\xi_{\mathsf{P}}\|_{\mathcal{B}(\mathsf{P})} = 1; \text{ with } \qquad \langle \xi_{\mathsf{P}}, \xi_{\mathsf{P},k} \rangle_{\mathcal{B}(\mathsf{P})} = 0,\, k < j, \quad (96)$$

over ξ_{P} in $\mathcal{B}^2(\mathsf{P})$, where $\langle f_{\mathsf{P},i}, \xi_{\mathsf{P}} \rangle_{\mathcal{B}(\mathsf{P})}$ is the projection of $f_{\mathsf{P},i}$ along the direction in $\mathcal{B}^2(\mathsf{P})$ identified by ξ_{P}, i.e., coordinate of f_{P} (Fourier coefficient). The orthogonality condition has only to be fulfilled for $j \geq 2$, and guarantees that the jth wSFPC $\xi_{\mathsf{P},j}$ is orthogonal to the first $j-1$ wSFPCs.

Since $\mathcal{B}^2(\mathsf{P})$ is a Hilbert space, the solution of the maximization problem (96) exists and is unique for all $j \in \{1, 2, \ldots, N-1\}$. It coincides with the set of eigenfunctions associated with the ordered eigenvalues of the sample covariance operator $V : \mathcal{B}^2(\mathsf{P}) \to \mathcal{B}^2(\mathsf{P})$, defined for $\xi_\mathsf{P} \in \mathcal{B}^2(\mathsf{P})$ as

$$V\xi_\mathsf{P} = \frac{1}{N} \odot_\mathsf{P} \bigoplus\nolimits_{\mathsf{P}_{i=1}}^{N} \langle f_{\mathsf{P},i}, \xi_\mathsf{P} \rangle_{\mathcal{B}(\mathsf{P})} \odot_\mathsf{P} f_{\mathsf{P},i}. \tag{97}$$

The jth wSFPC $\xi_{\mathsf{P},j}$ is thus obtained by solving the eigenequation $V\xi_{\mathsf{P},j} = \rho_j \odot_\mathsf{P} \xi_{\mathsf{P},j}$. The $N-1$ eigenvalues $\rho_1 \geq \ldots \geq \rho_{N-1}$ represent the variability of the dataset along the directions of the associated eigenfunctions $\xi_{\mathsf{P},1}, \ldots, \xi_{\mathsf{P},N-1}$.

From the practical viewpoint, it is desirable to restate the problem of finding the eigenpairs $(\xi_{\mathsf{P},j}, \rho_j), j = 1, \ldots, N-1$, in $\mathcal{B}^2(\mathsf{P})$ in terms of the unweighted L^2 spaces, i.e. $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$, where well-established computational methods are available. To this end, consider the clr_u transformation of the data, i.e. $\mathrm{clr}_u(f_{\mathsf{P},1}), \ldots, \mathrm{clr}_u(f_{\mathsf{P},N})$. Following the same arguments of [21], one can easily prove that performing a functional principal component analysis of the transformed dataset in $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ yields the eigenpairs $(\mathrm{clr}_u(\xi_{\mathsf{P},j}), \rho_j)\,, j = 1, \ldots, N-1$. The resulting eigenfunctions $\mathrm{clr}_u(\xi_{\mathsf{P},j})$ can be eventually transformed back into $\mathcal{B}^2(\mathsf{P})$, or into the unweighted $\mathcal{B}^2(\lambda)$, by using the corresponding inverse clr transformation (i.e. clr_u^{-1} or $\mathrm{clr}_{\sqrt{\mathsf{P}}}^{-1}$ respectively) to proceed with interpretation in the original space.

The results of wSFPCA can be interpreted, e.g. by analyzing the principal component scores, which are useful to inspect the relationships among observations. Note that the score f_{ij} is a projection of the (centered) observation $f_{\mathsf{P},i}$ along the direction $\xi_{\mathsf{P},j}$, i.e. $f_{ij} = \langle f_{\mathsf{P},i}, \xi_{\mathsf{P},j} \rangle_{\mathcal{B}(\mathsf{P})} = \langle \mathrm{clr}_u(f_{\mathsf{P},i}), \mathrm{clr}_u(\xi_{\mathsf{P},j}) \rangle_{L^2(\lambda)}$, and thus the scores coincide in $\mathcal{B}^2(\mathsf{P})$ and $L^2(\lambda)$. It is useful to visualize the mean density perturbed by the jth wSFPC $\xi_{\mathsf{P},j}$ powered by a suitable coefficient. This represents the variability around the mean function along the direction of a given wSFPC, and can support the analyst in the definition of a weighting strategy for the dataset at hand. Indeed, in the context of general reference measures, the wSFPCs can be plotted and interpreted to see the effect of weighting the domain of the distributional variable according to alternative reference measures. Finally, for the purpose of dimensionality reduction, the number of wSFPCs to be retained can be set by the commonly used scree plot. Particularly, searching for an elbow shape or setting a threshold on the portion of variance explained by wSFPCs as usually.

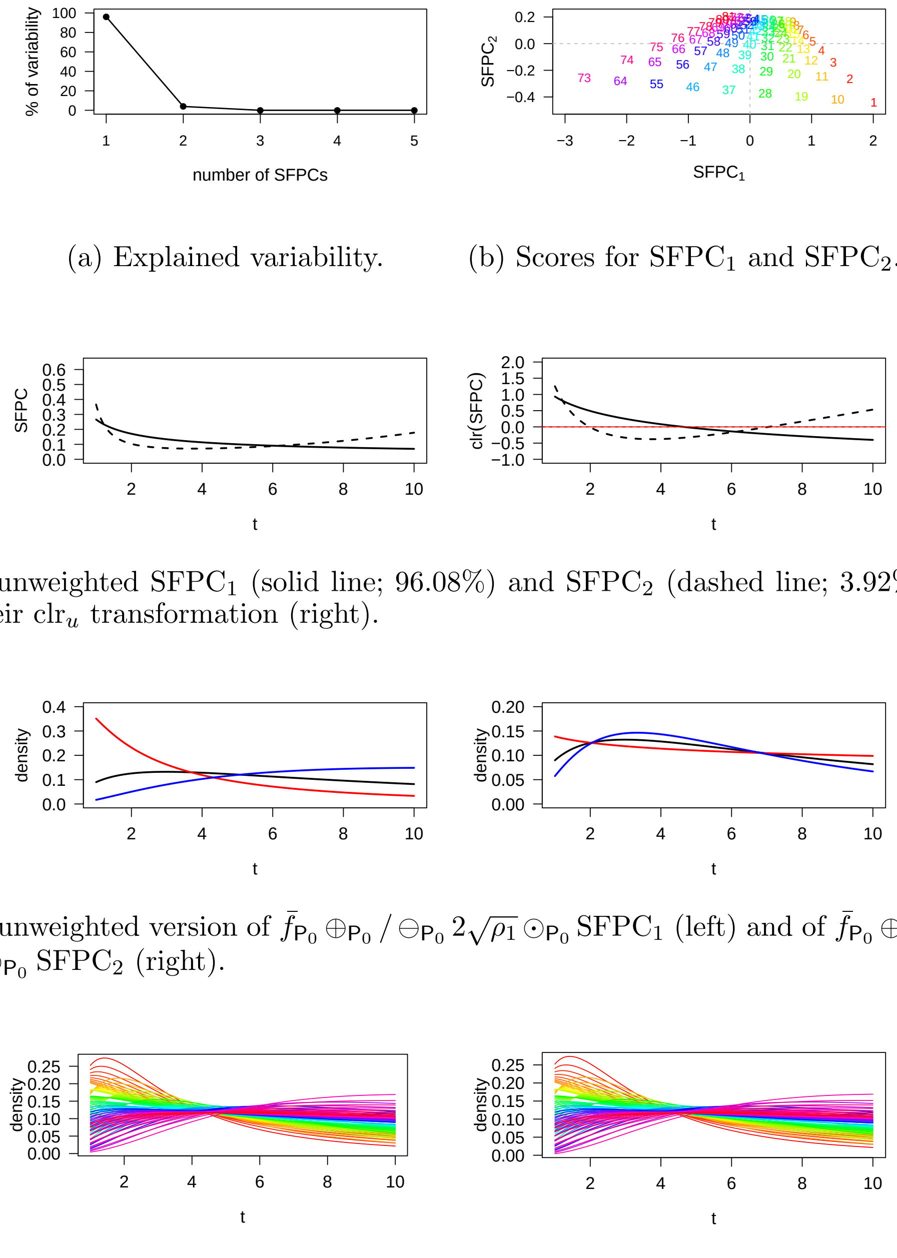
5.1 Changing the reference measure: consequences on SFPCA using simulated densities from exponential families

In this section, the effect of changing the reference measure is further analyzed in the context of weighted SFPCA. The same set of log-normal densities used in Section 4.3 is considered, by setting the reference measure to either uniform or exponential distribution.

Both datasets considered in Section 4.3 belong to a 2-parametric exponential family which forms an affine subspace of the Bayes space whose dimension is precisely the number of parameters [42]. This feature was highlighted in [21] for the case of the Lebesgue reference measure. Accordingly, the original spaces can be reconstructed (without loss of information) by the first two SFPCs (SFPC$_1$ and SFPC$_2$), forming an orthonormal basis of the corresponding affine subspace. One may expect that changing the reference measure for densities in the exponential family will have an impact on the wSFPCA while preserving the data dimensionality. We also note that the results of wSFPCA under a uniform reference measure are expected to be just a rescaling of those that would be obtained with the SFPCA of [21].

Figures 31 and 32 report the wSFPCA results on the log-normal densities when uniform and exponential reference measures are used respectively. As expected, the first two wSFPCs represent the total variability of the dataset in all cases.

When placing more emphasis on the left-hand side of the support Ω by increasing the parameter δ of the exponential reference measure (Figure 32), the portion of explained variability increases in SFPC$_1$ (and thus decreases in SFPC$_2$). Regardless of the reference measure, the first clr-wSFPC suggests that the main contribution to the total variability is associated with a contrast between the



(a) Explained variability.

(b) Scores for SFPC$_1$ and SFPC$_2$.

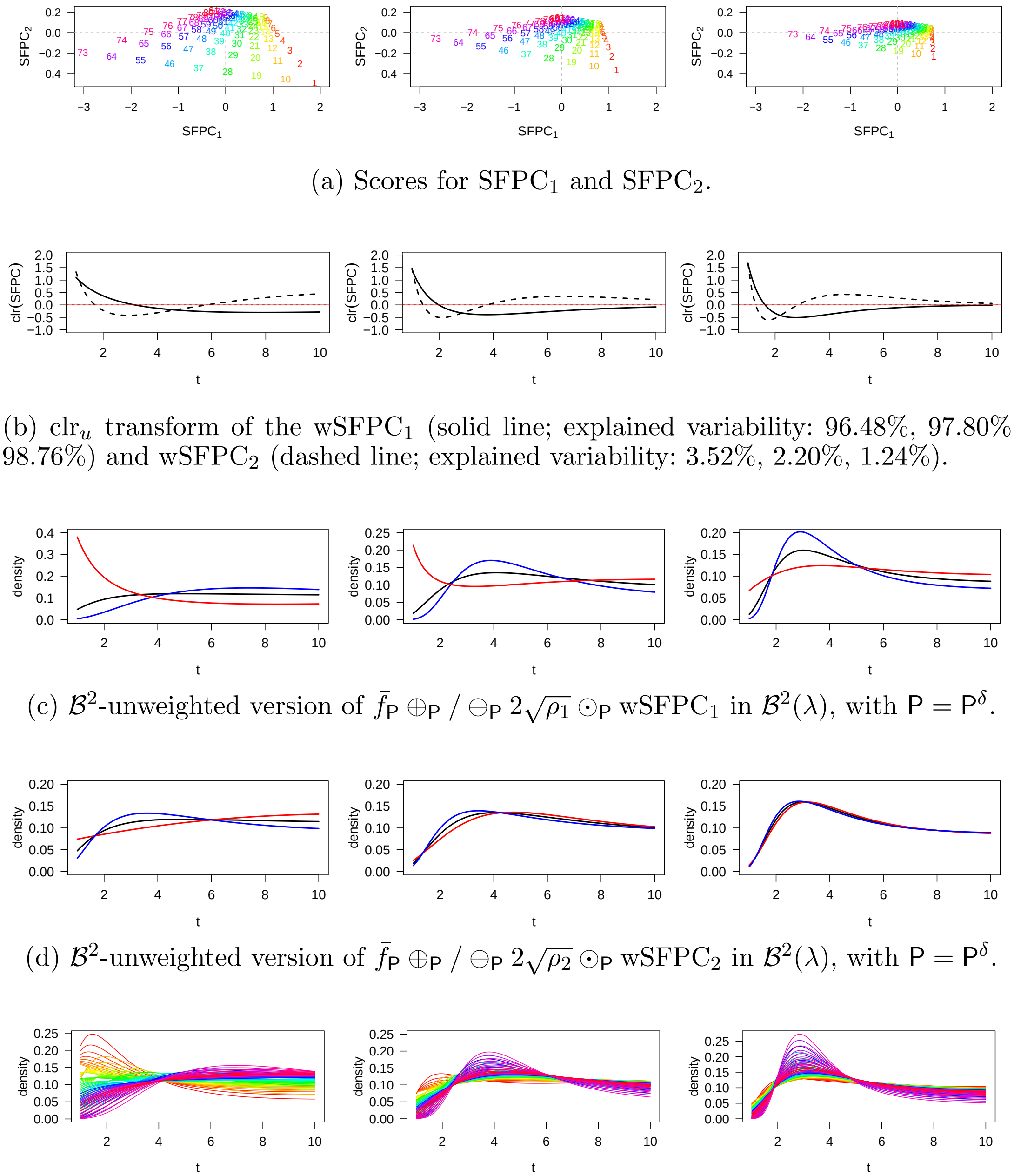
(c) $\mathcal{B}^2$-unweighted SFPC$_1$ (solid line; 96.08%) and SFPC$_2$ (dashed line; 3.92%) (left) and their clr$_u$ transformation (right).

(d) $\mathcal{B}^2$-unweighted version of $\bar{f}_{\mathsf{P}_0} \oplus_{\mathsf{P}_0} / \ominus_{\mathsf{P}_0} 2\sqrt{\rho_1} \odot_{\mathsf{P}_0}$ SFPC$_1$ (left) and of $\bar{f}_{\mathsf{P}_0} \oplus_{\mathsf{P}_0} / \ominus_{\mathsf{P}_0} 2\sqrt{\rho_2} \odot_{\mathsf{P}_0}$ SFPC$_2$ (right).

(e) $\mathcal{B}^2$-unweighted version of the P_0-densities $f_{\mathsf{P}_0,ij}$.

(f) $\mathcal{B}^2$-unweighted version of the approximation of $f_{\mathsf{P}_0,ij}$ via SFPC$_1$ and SFPC$_2$.

Figure 31: Results of SFPCA for simulated log-normal densities in the case of a uniform reference measure P_0. Results in panels (c) to (f) are represented in the unweighted spaces $L^2_{0,\sqrt{\mathsf{P}_0}}(\lambda)$, $\mathcal{B}^2(\lambda)$. By $\mathcal{B}^2$-unweighted version of $f \in \mathcal{B}^2(\mathsf{P}_0)$ it is meant $\omega^{-1}(f_{\mathsf{P}_0}) \in \mathcal{B}^2(\lambda)$.



(a) Scores for SFPC$_1$ and SFPC$_2$.

(b) clr$_u$ transform of the wSFPC$_1$ (solid line; explained variability: 96.48%, 97.80%, 98.76%) and wSFPC$_2$ (dashed line; explained variability: 3.52%, 2.20%, 1.24%).

(c) $\mathcal{B}^2$-unweighted version of $\bar{f}_{\mathsf{P}} \oplus_{\mathsf{P}} / \ominus_{\mathsf{P}} 2\sqrt{\rho_1} \odot_{\mathsf{P}}$ wSFPC$_1$ in $\mathcal{B}^2(\lambda)$, with $\mathsf{P} = \mathsf{P}^{\delta}$.

(d) $\mathcal{B}^2$-unweighted version of $\bar{f}_{\mathsf{P}} \oplus_{\mathsf{P}} / \ominus_{\mathsf{P}} 2\sqrt{\rho_2} \odot_{\mathsf{P}}$ wSFPC$_2$ in $\mathcal{B}^2(\lambda)$, with $\mathsf{P} = \mathsf{P}^{\delta}$.

(e) $\mathcal{B}^2$-unweighted version of the approximation of the P-densities via wSFPC$_1$ and wSFPC$_2$.

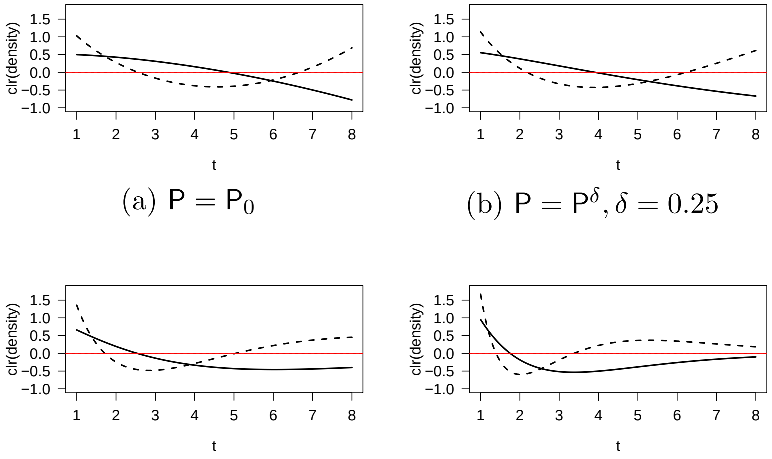
Figure 32: Results of SFPCA for simulated log-normal densities in case of exponential reference measures with $\delta = 0.25$ (first column), $\delta = 0.75$ (second column) and $\delta = 1.25$ (third column). By $\mathcal{B}^2$-unweighted version of $f \in \mathcal{B}^2(\mathsf{P}^{\delta})$ it is meant $\omega^{-1}(f_{\mathsf{P}^{\delta}}) \in \mathcal{B}^2(\lambda)$

left-hand side of the domain and the other side. It should be pointed out that clr transformed densities always display a contrast due to the zero integral constraint. However, it is worth noticing that the zero-crossing point moves to the left when the reference measure is changed using higher values of δ. The same pattern is observed for the second clr-wSFPC since it still highlights the variability in the left-hand side of the domain, but additionally it presents a contrast between the central and the rest. These conclusions are further supported by Figures 31d, 32c and 32d, where, for $\mathsf{P} = \mathsf{P}_0$ and $\mathsf{P} = \mathsf{P}^\delta$ respectively, the mean density is perturbed ($\oplus_\mathsf{P}/\ominus_\mathsf{P}$) by the SFPC powered ($\odot_\mathsf{P}$) to twice the standard deviation $\sqrt{\rho}$ along the corresponding direction ξ_P (i.e. $\bar{f}_\mathsf{P} \oplus_\mathsf{P} / \ominus_\mathsf{P} (2\sqrt{\rho_j} \odot_\mathsf{P} \xi_{\mathsf{P},j})$, where the $(\rho_j, \xi_{\mathsf{P},j})$ is the jth eigenpair of the covariance operator V). These results suggest that, when a uniform reference P_0 or an exponential P^δ with $\delta = 0.25$ are considered, the main mode of variability resides in the left-hand side of the domain. Changing the reference measure to P^δ has the effect of inflating the variability of the data in the central-left section of the domain (around the interval $[2, 4]$, see also Figure 27d), with a direct effect on the variability displayed along the first wSFPC.

Figures 31b and 32a display the score plots of wSFPCA under a P_0 and P^δ respectively. The symbols represent the indices of the data points, with $f_{\mathsf{P},ij}$ being represented through the index $\kappa = j + 9(i-1)$, $i, j = 1, ..., 9$. Recalling that the sampling design considers $\mu_i = 0.6 + 0.25 \cdot (i-1)$ and $\sigma_j = 0.5 + 0.07 \cdot (j-1)$ for $i, j = 1, \ldots, 9$; note that SFPC$_1$ arranges the densities according to parameter μ_i whereas SFPC$_2$ according to parameter σ_j.

Finally, Figures 31f and 32e display the projection of the log-normal densities on the basis generated by the first two wSFPCs, each represented in the unweighted $\mathcal{B}^2(\lambda)$ space (i.e., after clr$_u$ transformation). These results confirm that the dimensionality of the affine spaces of $\mathcal{B}^2(\mathsf{P})$, for $\mathsf{P} = \mathsf{P}_0$ and $\mathsf{P} = \mathsf{P}^\delta$, spanned by the log-normal family is indeed captured by the first two wSFPCs.

Consistent results are obtained for the second data set consisting of Weibull densities. For the sake of brevity, only evolution in exponentially weighted SFPCs is illustrated and compared with the unweighted case in Figure 33. The plots of the wSFPCs confirm that most of the relative variability is contained in the right tail of the distribution (uniform case) and this effect is getting suppressed when



(a) $\mathsf{P} = \mathsf{P}_0$

(b) $\mathsf{P} = \mathsf{P}^{\delta}, \delta = 0.25$

(c) $\mathsf{P} = \mathsf{P}^{\delta}, \delta = 0.75$

(d) $\mathsf{P} = \mathsf{P}^{\delta}, \delta = 1.50$

Figure 33: Results of wSFPCA for simulated Weibull densities in case of uniform and exponential reference measure with $\delta \in \{0.25, 0.75, 1.50\}$: clr_u transform of the wSFPC$_1$ (solid line; proportion of explained variability: 99.62, 99.45, 98.78, 97.82) and of the wSFPC$_2$ (dashed line; proportion of explained variability: 0.38, 0.55, 1.22, 2.18).

weighting is employed. Moreover, in all the cases, wSFPCA properly singles out just two non-zero wSFPCs. However, these two components are associated with different explained variance for different references, the wSFPC$_1$ decreasing and wSFPC$_2$ increasing for increasing values of δ.

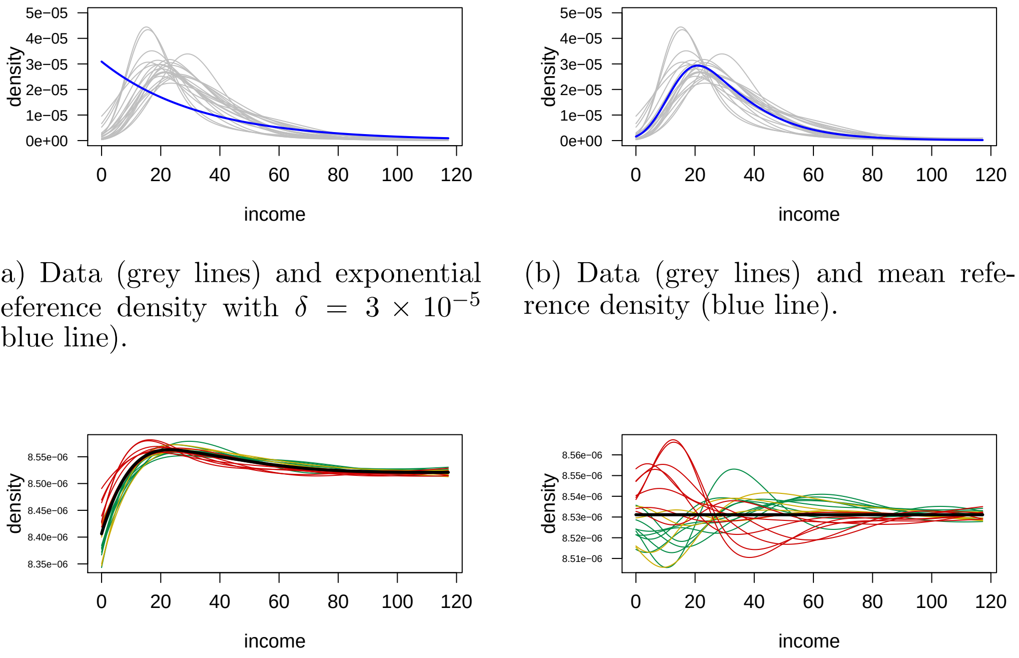
5.2 Application: weighted SFPCA of Italian income data

As an illustrative example, we apply wSFPCA to income data introduced in Section 2.2.3. In the following, we describe the results of wSFPCA when the reference measure is set to (i) the Lebesgue measure, (ii) the exponential measure P^{δ} (Section 4.3), and (iii) the measure P^{m} corresponding to the unweighted sample mean of the data as in [43]. Figure 34 displays the (ii) and (iii) cases, together with the corresponding $\mathcal{B}^2$-unweighted densities $(\omega^{-1}(f_{\mathsf{P}}))$. Finally, as (iv) we apply also Burr measure which is popularly used in economic studies.

SFPCA w.r.t. the Lebesgue measure SFPCA was performed by considering the Lebesgue reference measure as in [21]. The results are reported in the first column of Figure 35. Figure 35b displays the clr_λ transform of the first two SFPCs. The first clr-SFPC is interpreted as a contrast between the bottom band of the income distribution (i.e., income lower than 36.6 k€) and the rest. The second SFPC still contrasts low against high incomes, but provides further insight into differences in the central band of the distribution (i.e. for middle-income values). These findings are also well reflected in Figure 35c-d, which displays variation along the first and second SFPCs respectively with respect to the sample mean. Having fixed the sign of the clr-SFPC as in Figure 35b, high scores along the first principal direction are predominantly associated with regions characterized by more low-income households than the average and, conversely, low scores are expected for high-income regions. Similarly, Figure 35d supports the interpretation of the second principal direction. From Figure 35a, the first SFPC can be clearly associated with geographical location, as the northern and central regions (higher incomes) appear well separated from the southern regions (lower incomes) along this direction. Finally, the approximation of smoothed density data using only the first SFPC is shown in Figure 35e. Comparing this with the actual data (Figure 10b), the goodness of the approximation can be appreciated.

wSFPCA w.r.t. exponential measure An exponential reference measure $\mathsf{P} = \mathsf{P}^\delta$ was used in order to emphasize the relative scale of income values, with δ optimizing a data-driven criterion. In particular, δ maximizes regional discrimination along the first principal directions. We remark that other criteria may be of interest, e.g. one may want to attain a certain rate of explained variability by the first SFPCs, or to select the reference measure that best fits the data. Following our criterion, Table 4 presents the classic decomposition of the total sum of squares (SS_T) into between-groups (SS_B) and within-groups (SS_W) sum of squares when the scores for the wSFPC_1 using $\mathsf{P} = \mathsf{P}^\delta$ are modeled via a one-way analysis of variance (ANOVA), using the Italian regions (north, center, south) as a factor. Amongst the tested reference measures P^δ, we selected the one associated with the highest ratio $\mathrm{SS}_B/\mathrm{SS}_T$ (i.e. the highest discrimination between groups), which is $\delta = 3 \times 10^{-5}$. Note that we could otherwise consider Fisher's canonical direction as in ordinary discriminant analysis, which provides the direction of maximum discrimination between groups.

SFPCA was performed on the dataset consisting of exponentially weighted



(a) Data (grey lines) and exponential reference density with $\delta = 3 \times 10^{-5}$ (blue line).

(b) Data (grey lines) and mean reference density (blue line).

(c) $\mathcal{B}^2$-unweighted version of P-densities when $\mathsf{P} = \mathsf{P}^{\delta}$ and their mean function (black line).

(d) $\mathcal{B}^2$-unweighted version of P-densities when $\mathsf{P} = \mathsf{P}^{m}$ and their mean function (black line).

Figure 34: Income densities in case of exponential (left column) and mean reference measure (right column). By $\mathcal{B}^2$-unweighted version of $f \in \mathcal{B}^2(\mathsf{P})$ it is meant $\omega^{-1}(f_{\mathsf{P}}) \in \mathcal{B}^2(\lambda)$.

distributions (Figures 34a-c). The results are reported in the second column of Figure 35. The score plot (Figure 35a) shows that the configuration of the scores well represents geographical locations, even though it is somehow similar to the one obtained obtain using the Lebesgue reference measure. However, the amount of variability explained along the first two SFPCs is higher in comparison to the unweighted case (Figure 35b).

It is worth noticing that the interpretation of the wSFPCs appears to be affected by the change in the reference measure. Indeed, although the first SFPC (Figure 35b) still represents a contrast between low and high incomes households, the second SFPC displays a contrast within the low-income group. This could prompt further interesting economic interpretation, as this contrast might be related to an unequal redistribution of wealth within the lower-income class. For

	SS$_B$	SS$_W$	SS$_T$	SS$_B$/SS$_T$
Uniform	1.5030	0.7130	2.2160	0.6782
Exp(1.5×10^{-5})	2.1811	0.8425	3.0236	0.7214
Exp(3×10^{-5})	2.3631	0.9111	3.2742	**0.7217**
Exp(6×10^{-5})	1.8540	0.9128	2.7668	0.6701
Exp(1.2×10^{-4})	0.8609	0.8515	1.7124	0.5027

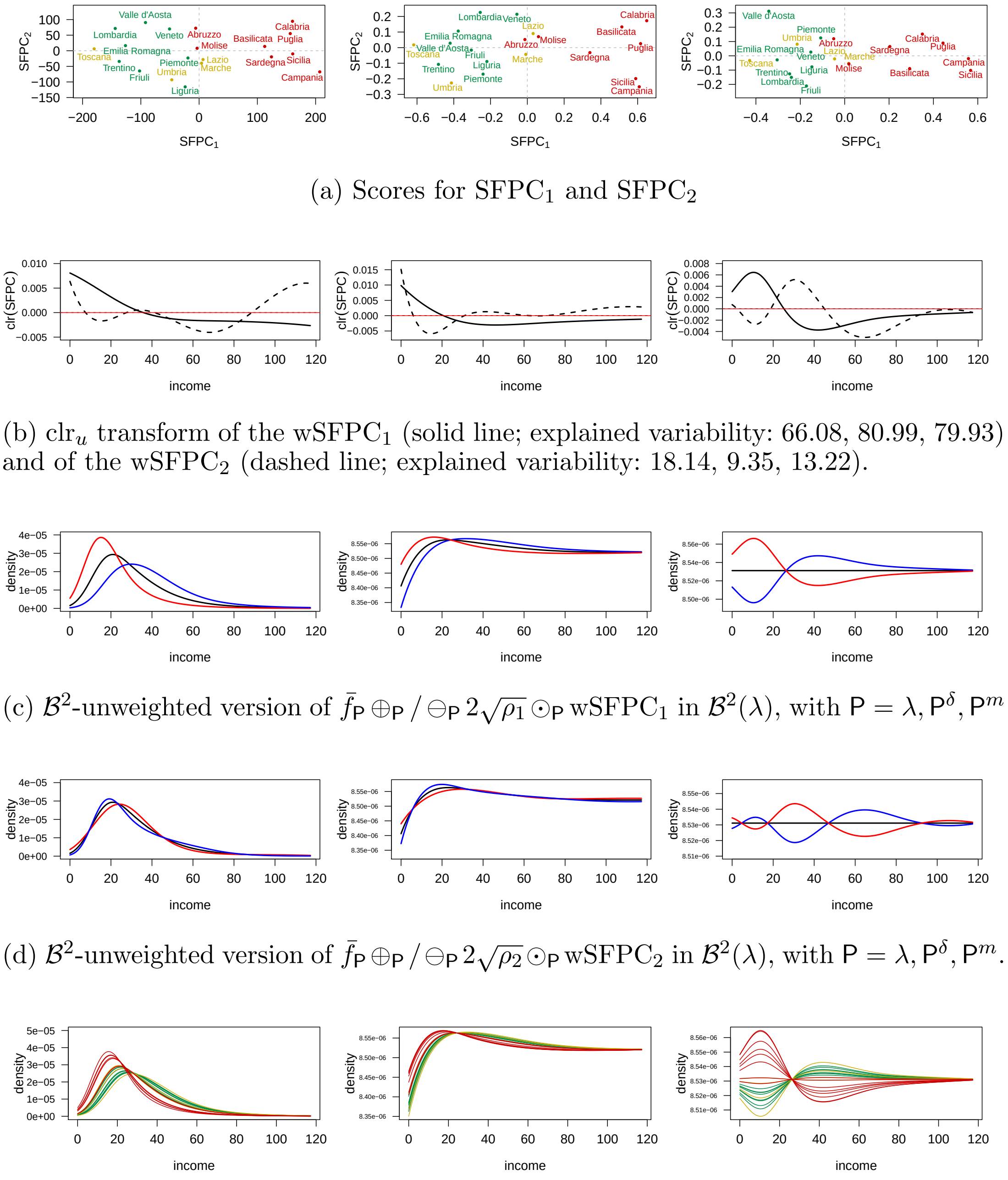
Table 4: ANOVA sum of squares decomposition for the first SFPC scores based on uniform and exponential reference measures using region as factor. The exponential measures were P^{δ}, with $\delta \in \{1.5 \times 10^{-5}, 3 \times 10^{-5}, 6 \times 10^{-5}, 1.2 \times 10^{-4}\}$.

instance, the (annual) poverty threshold in 2008 for a household of two members was 11.8 k€, which is roughly half of the zero-crossing of SFPC$_1$ and close to the zero-crossing of SFPC$_2$. Hence, weighting according to the relative scale of income data could help to signaling unequal redistribution of wealth, particularly amongst the low-income population.

wSFPCA w.r.t. the sample mean A different view is obtained when the reference measure is set to the sample mean of the data $\bar{f}$ (density w.r.t. the Lebesgue reference measure), computed as

$$\bar{f}(t) = \frac{1}{N} \odot \bigoplus_{i=1}^{N} f_i(t), \qquad t \in \Omega.$$

Recall that the reference measure determines the origin of the space $\mathcal{B}^2(\mathsf{P})$, which is a P-density represented by a constant function. This is unchanged when mapped to $\mathcal{B}^2(\lambda)$ through the $\mathcal{B}^2$-unweighting map ω^{-1}. For this reason, the sample weighted mean density in $\mathcal{B}^2(\lambda)$ appears as an uniform density in Figure 34d. In this case, the representation of the $\mathcal{B}^2$-unweighted data (Figure 34d) provides additional information about the dispersion of income distributions around their mean. Note that this has to be interpreted as usual (unweighted) PDFs. The distributions vary in different ways across regions: the income distributions in southern regions tend to be more concentrated than the average around low income (the average being represented as a uniform distribution), and they are less concentrated for higher incomes. The opposite is observed for northern and central regions. This is also well reflected by the wSFPCA output which is summarized in the third column of Figure 35.



(a) Scores for SFPC$_1$ and SFPC$_2$

(b) clr$_u$ transform of the wSFPC$_1$ (solid line; explained variability: 66.08, 80.99, 79.93) and of the wSFPC$_2$ (dashed line; explained variability: 18.14, 9.35, 13.22).

(c) $\mathcal{B}^2$-unweighted version of $\bar{f}_{\mathsf{P}} \oplus_{\mathsf{P}} / \ominus_{\mathsf{P}} 2\sqrt{\rho_1} \odot_{\mathsf{P}}$ wSFPC$_1$ in $\mathcal{B}^2(\lambda)$, with $\mathsf{P} = \lambda, \mathsf{P}^\delta, \mathsf{P}^m$

(d) $\mathcal{B}^2$-unweighted version of $\bar{f}_{\mathsf{P}} \oplus_{\mathsf{P}} / \ominus_{\mathsf{P}} 2\sqrt{\rho_2} \odot_{\mathsf{P}}$ wSFPC$_2$ in $\mathcal{B}^2(\lambda)$, with $\mathsf{P} = \lambda, \mathsf{P}^\delta, \mathsf{P}^m$.

(e) $\mathcal{B}^2$-unweighted version of the approximation of the P-densities via SFPC$_1$.

Figure 35: SFPCA results for income densities in Italian regions in case of reference measure set to (1) $\mathsf{P} = \lambda$ the Lebesgue measure (first column), (2) $\mathsf{P} = \mathsf{P}^\delta$ the exponential measure with $\delta = 3 \times 10^{-5}$ (second column) and (3) $\mathsf{P} = \mathsf{P}^m$ the mean measure (third column). By $\mathcal{B}^2$-unweighted version of $f \in \mathcal{B}^2(\mathsf{P})$ it is meant $\omega^{-1}(f_{\mathsf{P}}) \in \mathcal{B}^2(\lambda)$.

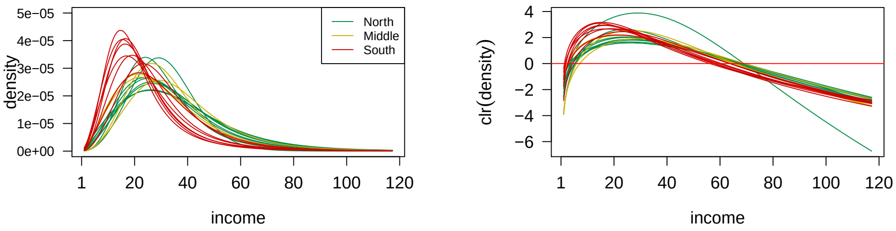
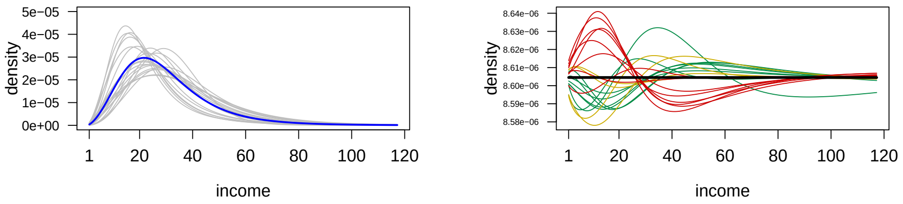


Figure 36: Fitted income densities by Burr (Singh-Maddala) distribution in $\mathcal{B}^2(\lambda)$ (left) space and its clr transformation in $L_0^2(\lambda)$ (right) space with the same color resolution as in the map in Figure 8; income is expressed in 10^3 k€.

The first wSFPC – depicted in panel (b) – still contrasts the bottom of the distributions (income below 25.23 k€) against their middle and top. The second wSFPC shows differences especially between the middle band of the distributions (income in $[18.77, 45.06]$ k€) and the top band (income over 45.06 k€). Note that a higher dispersion of the scores wSFPC$_2$ is observed for northern and central regions in relation to southern regions, which appear almost constant along the second mode of variation. In fact, wSFPC$_2$ seems to reveal a different distribution of wealth in the central band of the income distributions. Lombardia and Friuli regions tend to concentrate more medium-high incomes than the mean. Contrarily, the Valle d'Aosta region is characterized by low-medium incomes, appearing as an outlier along wSFPC$_2$. The approximation of the sampled income distributions by the first SFPC (capturing almost 80% of the total variability) well reflects the data structure, as can be appreciated by comparing Figure 34d and Figure 35e.

wSFPCA w.r.t. a Burr reference measure In economic studies, parametric distributions are often used to represent income data. For instance, two-parameter distributions defined on positive real line such as log-normal and Weibull distributions are commonly used for this purpose [27]. We here fit the raw income data by a three-parameter distribution known as Burr (Singh-Maddala) using the theory of generalized linear models (implemented with a help of R-package VGAM; function `vglm`), and having the following density

$$f_\lambda(t;\alpha,c,k) = \frac{k\cdot c}{\alpha}\left(\frac{t}{\alpha}\right)^{c-1}\left[1+\left(\frac{t}{\alpha}\right)^{c}\right]^{-k-1}, \quad t>0, \alpha>0, c>0, k>0,$$



(a) Data (grey lines) and mean reference density (blue line).

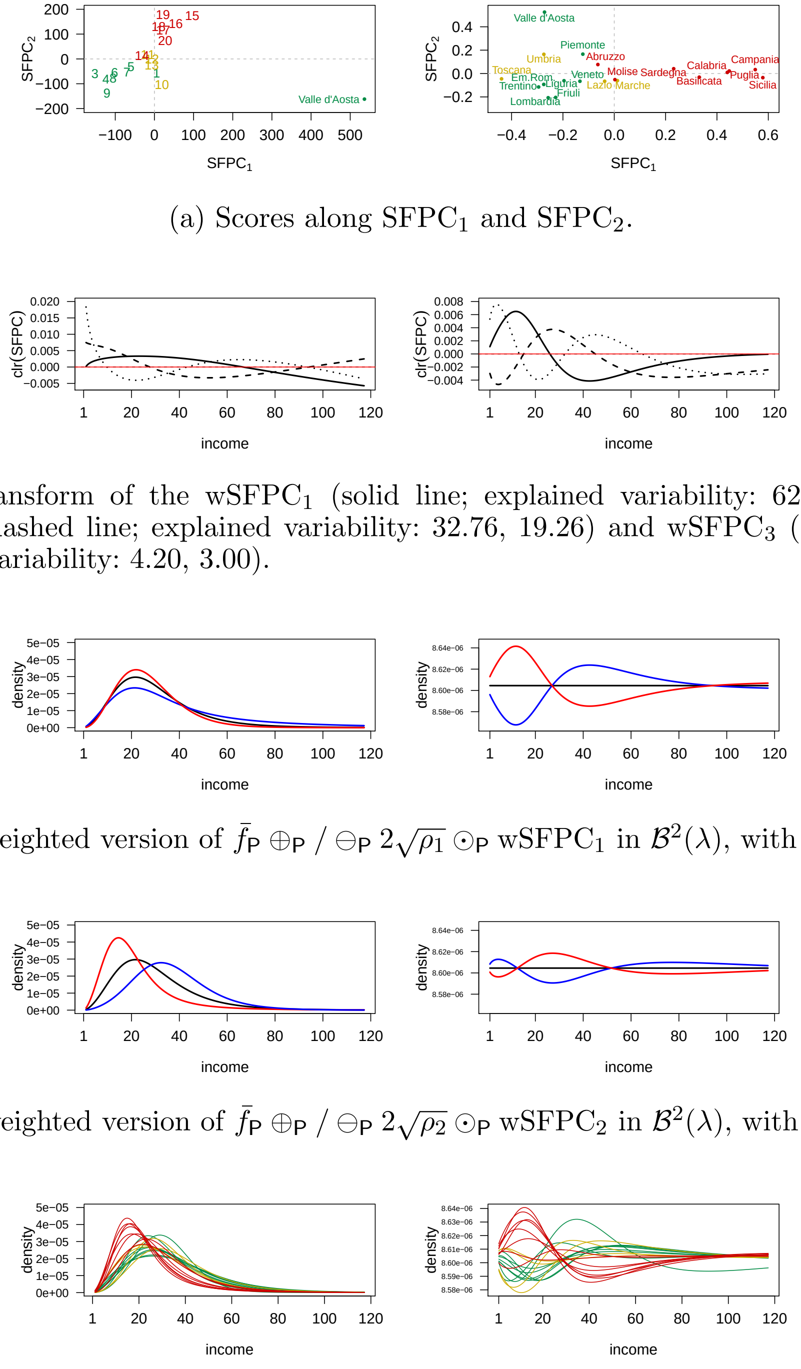
(b) $\mathcal{B}^2$-unweighted version of P-densities when $\mathsf{P} = \mathsf{P}^m$ and their mean function (black line).

Figure 37: Burr (Singh-Maddala) income densities in case of mean reference measure. By $\mathcal{B}^2$-unweighted version of $f \in \mathcal{B}^2(\mathsf{P})$ it is meant $\omega^{-1}(f_{\mathsf{P}}) \in \mathcal{B}^2(\lambda)$.

where c and k are the shape parameters and α is the scale parameter. We remark that all the above-mentioned distributions come from the exponential family. As such, the dimension of the corresponding space can be correctly determined using the Bayes space geometry and it corresponds to the number of parameters. The raw income data were fitted on the truncated interval $\Omega = [1, 117.218]$ k€ in order to avoid artifacts related to very low values of the fitted densities on the left boundary of the domain Ω that would bias the subsequent statistical processing. Figure 36 shows the fitted data. The graphical inspection of their clr representation clearly shows that this approach might not be appropriate for all observations, the region Valle d'Aosta clearly representing an outlier. Accordingly, the respective region (Valle d'Aosta) could be possibly removed from further analysis if needed.

Finally, SFPCA is performed for all fitted densities for the cases of the Lebesgue measure and Burr (Singh-Maddala) mean distribution P^m (computed from fitted data, see Figure 37), respectively. The results of SFPCA for these data sets are reported in Figure 38. It can be noted first that negligible information is lost since the cumulative variability captured by the first three SFPCs equals 99.85% and 99.86%, respectively (due to possible numerical imprecision, not 100% was achieved). The data representation did not influence the conclusions for the mean reference (see also the third column of Figure 35), but this is no longer true if we compare results for in the case of the Lebesgue measure (see also the first column of Figure 35). The outlier Valle d'Aosta region, identified previously, is further confirmed from the scree plot in panel (a), and contributes substantially to variability in the right part of the distribution (see a plot of SFPC_1 in panel (b)). However, one may appreciate that working with distributions from exponential family guarantee the dimension of data. Indeed, the exact reconstruction of data can be inspected from comparison of Figures 36 (left) and 37 (right) with 38e.

Both the exponential and sample mean reference measures used above can be considered as data-driven. The former is particularly suitable for income densities, because it reflects the relative scale of income data, and the parameter of the exponential density can be determined by optimizing a data-driven criterion – here set in the framework of SFPCA (see Table 4). However, the sample mean $\bar{f}(t)$ itself may be more appropriate as a default choice of the reference measure. In this case, the input data are directly used to drive the choice of the reference, highlighting departures from the mean trend in the data. A more general setting $\bar{f}(t)^{\alpha}$ (possibly rescaled to a given integral constraint representation to achieve comparability) can be also considered, where α is a real parameter to be set by the analyst. Here, setting the parameter α in $[0, 1]$ would lead to a balance between the extreme cases of $\alpha = 0$ (uniform reference) and $\alpha = 1$ (sample mean). For instance, the "compromise" choice $\alpha = 0.5$ would allow focusing on the main trend while down-weighting subdomains with lower density values on average. This choice would also reflect a similar strategy proposed in the multivariate case of compositional data [20], where smaller values of components are more likely to be burdened by relative scale effects, uncertainty and, in the functional context, also by the possible presence of count zeros [25] in the aggregated histogram data.



(a) Scores along SFPC$_1$ and SFPC$_2$.

(b) clr$_u$ transform of the wSFPC$_1$ (solid line; explained variability: 62.88, 77.58), wSFPC$_2$ (dashed line; explained variability: 32.76, 19.26) and wSFPC$_3$ (dotted line; explained variability: 4.20, 3.00).

(c) $\mathcal{B}^2$-unweighted version of $\bar{f}_\mathsf{P} \oplus_\mathsf{P} / \ominus_\mathsf{P} 2\sqrt{\rho_1} \odot_\mathsf{P}$ wSFPC$_1$ in $\mathcal{B}^2(\lambda)$, with $\mathsf{P} = \lambda, \mathsf{P}^m$.

(d) $\mathcal{B}^2$-unweighted version of $\bar{f}_\mathsf{P} \oplus_\mathsf{P} / \ominus_\mathsf{P} 2\sqrt{\rho_2} \odot_\mathsf{P}$ wSFPC$_2$ in $\mathcal{B}^2(\lambda)$, with $\mathsf{P} = \lambda, \mathsf{P}^m$.

(e) $\mathcal{B}^2$-unweighted version of the approximation of the P-densities via SFPC$_1$, SFPC$_2$ and SFPC$_3$.

Figure 38: SFPCA results for fitted income densities in Italian regions by Burr (Singh-Maddala) distribution in case of reference measure set to (1) $\mathsf{P} = \lambda$ the Lebesgue measure (first column) and (2) $\mathsf{P} = \mathsf{P}^m$ the mean measure (second column). By $\mathcal{B}^2$-unweighted version of $f \in \mathcal{B}^2(\mathsf{P})$ it is meant $\omega^{-1}(f_\mathsf{P}) \in \mathcal{B}^2(\lambda)$. In panel (a), the same numbering of the regions is used as in the map in Figure 8.

Conclusions

The focus of this work was to develop statistical methods for the analysis of functional data carrying relative information – probability density functions, defined on a bounded domain. These methods are grounded on the theory of Bayes Hilbert spaces, capturing all key inherent features of densities (i.e., scale invariance, relative scale), and they extend the well-known results of FDA to density functions.

In Section 2, we considered the problem of statistical preprocessing of densities using spline functions, performed in the clr space. Firstly, we recalled optimal smoothing splines for clr transformed density functions as proposed in [23]. Here, we proved a new key result to characterize B-spline representation of clr transformed densities using standard B-spline basis system in terms of a linear constraint on the B-spline basis coefficients. Nevertheless, it was recognized that using the standard B-spline basis system for approximation of density functions in clr space has some limitations since the basis elements do not belong to the L_0^2 space. Therefore, this approach was updated by proposing a new class of compositional splines which enable to construct a B-spline basis directly in the clr space of density functions (ZB-spline basis system) and, consequently, also in the original space of densities (CB-spline basis system). Accordingly, compositional splines can be implemented instead of the standard ones into FDA methods for statistical processing of density functions. Also further tuning of the compositional splines is possible, here represented by the smoothing compositional splines or by orthonormalization of the ZB-basis systems. As for future research, it could be attractive to generalize the methodology of compositional splines even for multidimensional density functions.

In Section 3, a novel approach to perform functional regression when the response is a density function using the Bayes space methodology was developed. For the actual estimation of the regression coefficients, an approach based on B-spline expansion of clr transformed density functions was proposed. This expansion enables to control the smoothness of the estimated regression coefficients (density functions) through the smoothness of the B-spline representation of the response. On the other hand, it turned out that the linear constraint on B-spline basis coefficients (using the former approach for B-spline expansion of

PFDs) induces the singularity problem into the regression model. Nevertheless, this can be overcome by using the compositional splines which lead to expression of PDFs through a set of unconstrained coefficients. Such representation can be then further used for the purpose of inference on the coefficients using proper functional tests.

The role of reference measure in Bayes spaces was discussed in Section 4, specially, a novel weighting approach to probability density functions was proposed. An advanced weighting scheme was developed which enables to link weighted Bayes spaces to unweighted $\mathcal{B}^2$ and the L^2 spaces. The advantage of representing weighted densities in an unweighted space is demonstrated by the possibility of (i) making comparisons of densities arising from different weighting criteria, and (ii) visually interpret the results through ordinary 'unweighted eyes'. In fact, the proposed framework allows to perform statistical processing in weighted Bayes spaces by using simply popular (unweighted) methods, which were developed for FDA. In the final Section 5, this strategy has been demonstrated by extending a dimensionality reduction method (SFPCA) to the weighted case. Nevertheless, other methods could be considered as well, such as clustering, regression, spatial prediction techniques, etc. We finally stress that considering different weighting schemes can be particularly relevant in statistical applications, i.e., (i) to account for different degrees of uncertainty across the domain of the data, (ii) to incorporate prior knowledge about the phenomenon or (iii) to perform domain selection.

I truly hope that the presented thesis helps to expand the Bayes space methodology for statistical processing of density functions and that it will be a motivation to propose other statistical methods for analyzing PDFs such as outlier detection and related anomaly detection, classification or functional regression with densities playing the role of the response and/or covariates.

Bibliography

[1] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177, 1982.

[2] Bank of Italy. Survey on household income and wealth (SHIW). Available at https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/distribuzione-microdati/index.html?page=2, 2008.

[3] D. Billheimer, P. Guttorp, and W. Fagan. Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456):1205–1214, 2001.

[4] R. S. Bivand, J. Wilk, and T. Kossowski. Spatial association of population pyramids across Europe: The application of symbolic data, cluster analysis and join-count tests. *Spatial Statistics*, 21:339–361, 2017.

[5] H. Chen, P. T. Reiss, and T. Tarpey. Optimally weighted l2 distance for functional data. *Biometrics*, 70(3):516–525, 2014.

[6] C. De Boor. *A practical guide to splines*, Springer-Verlag, New York, 1978.

[7] P. Delicado. Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis*, 55(1):401–420, 2011.

[8] P. Dierckx. *Curve and surface fitting with splines.* Oxford University Press, Oxford, 1995.

[9] J. J. Egozcue, V. Pawlowsky-Glahn, R. Tolosana-Delgado, M. Ortego, and K. van den Boogaart. Bayes spaces: use of improper distributions and exponential families. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 107(2):475–486, 2013.

[10] J. J. Egozcue, C. Barcelo-Vidal, J. Martín-Fernández, E. Jarauta-Bragulat, J. Díaz-Barrero, G. Mateu-Figueras, V. Pawlowsky-Glahn, and A. Buccianti. Elements of simplicial linear algebra and geometry. *Compositional data analysis: Theory and applications*, 139–157, 2011.

[11] J. J. Egozcue, P. Daunis-i Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser. Simplicial regression. The normal model. *Journal of Applied Probability and Statistics (JAPS)*, 6:87–108, 2012.

[12] J. J. Egozcue, J. L. Díaz-Barrero, and V. Pawlowsky-Glahn. Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*, 22(4):1175–1182, 2006.

[13] J. J. Egozcue and V. Pawlowsky-Glahn. Changing the reference measure in the simplex and its weighting effects. *Austrian Journal of Statistics*, 45(4):25–44, 2016.

[14] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[15] J. J. Faraway. Regression analysis for a functional response. *Technometrics*, 39(3):254–261, 1997.

[16] V. Ferro and S. Mirabile. Comparing particle size distribution analysis by sedimentation and laser diffraction method. *Journal of Agricultural Engineering*, 2:35–43, 2009.

[17] E. Fišerová, L. Kubáček, and P. Kunderová. *Linear Statistical Models: Regularity and Singularities*. Academia, 2007.

[18] D. Floriello and V. Vitelli. Sparse clustering of functional data. *Journal of Multivariate Analysis*, 154:1–18, 2017.

[19] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[20] M. Greenacre and P. Lewi. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of Classification*, 26(1):29–54, 2009.

[21] K. Hron, A. Menafoglio, M. Templ, K. Hruzová, and P. Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, 94:330–350, 2016.

[22] P. Kokoszka, H. Miao, A. Petersen, and H. L. Shang. Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting*, 35(4):1304–1317, 2019.

[23] J. Machalová, K. Hron, and G. S. Monti. Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*, 43(8):1419–1435, 2016.

[24] J. Machalová, R. Talská, K. Hron, and A. Gába. Compositional splines for representation of density functions. *Under review*.

[25] J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, 2015.

[26] G. Mateu-Figueras and V. Pawlowsky-Glahn. A critical approach to probability laws in geochemistry. *Math Geosci*, 40:489–502, 2008.

[27] J. B. McDonald and Y. J. Xu. A generalization of the beta distribution with applications. *Journal of Econometrics*, 66(1-2):133–152, 1995.

[28] A. Menafoglio, G. Gaetani, and P. Secchi. Random domain decompositions for object-oriented Kriging over complex domains. *Stochastic Environmental Research and Risk Assessment*, 32(12):3421–3437, 2018.

[29] A. Menafoglio, A. Guadagnini, and P. Secchi. A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment*, 28(7):1835–1851, 2014.

[30] A. Menafoglio, A. Guadagnini, and P. Secchi. Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a Bayes space approach. *Water Resources Research*, 52(8):5708–5726, 2016.

[31] A. Menafoglio, P. Secchi, and A. Guadagnini. A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. *Mathematical Geosciences*, 48(4):463–485, 2016.

[32] V. Pawlowsky-Glahn and J. J. Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(456):384–398, 2001.

[33] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. *Modeling and analysis of compositional data.* John Wiley & Sons, 2015.

[34] A. Petersen and H.-G. Müller. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1):183–218, 2016.

[35] A. Petersen and H.-G. Müller. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2):339–351, 2019.

[36] J. Ramsay and B. Silverman. Functional data analysis, second edition. *Springer Series in Statistics*, 2005.

[37] W.-K. Seo and B. Beare. Cointegrated linear processes in Bayes Hilbert space. *Statistics & Probability Letters*, 147:90–95, 2018.

[38] Q. Shen and H. Xu. Diagnostics for linear models with functional responses. *Technometrics*, 49(1):26–33, 2007.

[39] H. A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

[40] R. Talská, A. Menafoglio, K. Hron, J. J. Egozcue, and J. Palarea-Albaladejo. Weighting the domain of probability densities in functional data analysis. *Stat, accepted for publication.*

[41] R. Talská, A. Menafoglio, J. Machalová, K. Hron, and E. Fišerová. Compositional regression with functional response. *Computational Statistics & Data Analysis*, 123:66–85, 2018.

[42] K. G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes linear spaces. *SORT*, 34(4):201–222, 2010.

[43] K. G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194, 2014.

[44] K. Yosida. *Functional analysis.* Springer, Berlin, 1980.

Appendix A: Proof of Theorem 2.1

In the following the notation $s_k^{\mathbf{b}}(t)$ is used to emphasize the dependency on vector $\mathbf{b} = (b_{-k}, \ldots, b_g)'$. It is known that

$$\int\limits_a^b s_k^{\mathbf{b}}(t)\, dt \;=\; \left[s_{k+1}^{\mathbf{c}}(t)\right]_a^b,$$

for a vector $\mathbf{c}$, that is

$$s_k^{\mathbf{b}}(t) = \sum_{i=-k}^{g} b_i B_i^{k+1}(t) \;=\; \frac{d}{dt} \sum_{i=-k-1}^{g} c_i B_i^{k+2}(t) \;=\; \frac{d}{dt} s_{k+1}^{\mathbf{c}}(t). \tag{98}$$

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

$$d_i = \frac{k+1}{\lambda_{i+k+1} - \lambda_i}, \quad i = -k, \ldots, g; \tag{99}$$

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

$$c_g \;=\; \frac{b_g}{d_g} + \cdots + \frac{b_{-k}}{d_{-k}} + c_{-k-1}. \tag{100}$$

With respect to (98) it is evident that

$$\int\limits_a^b s_k^{\mathbf{b}}(t)\, dt \;=\; \left[s_{k+1}^{\mathbf{c}}(t)\right]_a^b \;=\; s_{k+1}^{\mathbf{c}}(\lambda_{g+1}) - s_{k+1}^{\mathbf{c}}(\lambda_0), \tag{101}$$

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

$$s_{k+1}^{\mathbf{c}}(\lambda_{g+1}) - s_{k+1}^{\mathbf{c}}(\lambda_0) = c_g - c_{-k-1}. \tag{102}$$

Thus

$$\int\limits_a^b s_k^{\mathbf{b}}(t)\,dt \;=\; c_g - c_{-k-1}. \tag{103}$$

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

$$\int\limits_a^b s_k^{\mathbf{b}}(t)\,dt \;=\; 0$$

is fulfilled if and only if

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

From (100) 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 (99) we easily get

$$\int\limits_a^b s_k^{\mathbf{b}}(t)\,dt \;=\; 0 \qquad \Leftrightarrow \qquad \sum_{i=-k}^{g} b_i\left(\lambda_{i+k+1} - \lambda_i\right) \;=\; 0.$$

Appendix B: Aggregated Italian income data

Region	loc.	Proportions of income classes, $\mathbf{W}_i = (W_{i1}, \ldots, W_{iD})', i = 1, \ldots, N$								
Piemonte	N	0.064	0.387	0.326	0.134	0.052	0.022	0.009	0.005	0.003
Valle d'Aosta	N	0.018	0.375	0.312	0.208	0.042	0.014	0.010	0.014	0.007
Lombardia	N	0.089	0.274	0.271	0.150	0.107	0.057	0.022	0.018	0.012
Trentino	N	0.052	0.320	0.285	0.128	0.134	0.029	0.041	0.006	0.006
Veneto	N	0.098	0.331	0.257	0.177	0.081	0.024	0.015	0.010	0.007
Friuli	N	0.084	0.320	0.232	0.168	0.088	0.068	0.028	0.008	0.004
Liguria	N	0.078	0.352	0.217	0.217	0.081	0.026	0.026	0.003	0.002
Emilia Romagna	N	0.062	0.303	0.278	0.189	0.085	0.045	0.017	0.016	0.006
Toscana	M	0.042	0.281	0.293	0.190	0.106	0.052	0.015	0.015	0.007
Umbria	M	0.052	0.351	0.337	0.157	0.056	0.026	0.015	0.004	0.002
Marche	M	0.115	0.401	0.219	0.150	0.061	0.032	0.014	0.006	0.003
Lazio	M	0.115	0.354	0.260	0.150	0.066	0.032	0.012	0.007	0.002
Abruzzo	S	0.100	0.364	0.299	0.144	0.045	0.030	0.004	0.010	0.005
Molise	S	0.124	0.357	0.277	0.109	0.080	0.022	0.022	0.005	0.003
Campania	S	0.238	0.483	0.169	0.066	0.019	0.016	0.006	0.002	0.001
Puglia	S	0.238	0.441	0.197	0.072	0.025	0.007	0.014	0.005	0.002
Basilicata	S	0.247	0.385	0.170	0.116	0.039	0.031	0.006	0.005	0.002
Calabria	S	0.230	0.419	0.209	0.078	0.042	0.005	0.010	0.003	0.003
Sicilia	S	0.247	0.476	0.165	0.053	0.029	0.014	0.012	0.002	0.002
Sardegna	S	0.167	0.425	0.217	0.123	0.044	0.015	0.006	0.003	0.002
Midpoints of intervals t_{ij}		6574	19591	32608	45625	58641	71658	84675	97692	110709

Table 5: Proportions of $D = 9$ income classes in $N = 20$ Italian regions. The values $t_{ij}, i = 1, \ldots, N, j = 1, \ldots, D$ are the midpoints of the income subintervals of $\Omega = [0, 117.22]$ k€.

Region	loc.	Raw density values, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iD})', i = 1, \ldots, N$								
Piemonte	N	4.90e-06	2.97e-05	2.50e-05	1.03e-05	4.00e-06	1.7e-06	7.0e-07	4.0e-07	2e-07
Valle d'Aosta	N	1.40e-06	2.88e-05	2.40e-05	1.60e-05	3.20e-06	1.1e-06	8.0e-07	1.1e-06	5e-07
Lombardia	N	6.80e-06	2.10e-05	2.08e-05	1.15e-05	8.20e-06	4.4e-06	1.7e-06	1.4e-06	9e-07
Trentino	N	4.00e-06	2.46e-05	2.19e-05	9.80e-06	1.03e-05	2.2e-06	3.1e-06	4.0e-07	4e-07
Veneto	N	7.50e-06	2.54e-05	1.97e-05	1.36e-05	6.20e-06	1.8e-06	1.2e-06	8.0e-07	5e-07
Friuli	N	6.40e-06	2.46e-05	1.78e-05	1.29e-05	6.80e-06	5.2e-06	2.1e-06	6.0e-07	3e-07
Liguria	N	6.00e-06	2.71e-05	1.66e-05	1.66e-05	6.20e-06	2.0e-06	2.0e-06	2.0e-07	1e-07
Emilia Romagna	N	4.80e-06	2.33e-05	2.13e-05	1.45e-05	6.50e-06	3.5e-06	1.3e-06	1.2e-06	4e-07
Toscana	M	3.20e-06	2.16e-05	2.25e-05	1.46e-05	8.20e-06	4.0e-06	1.1e-06	1.1e-06	5e-07
Umbria	M	4.00e-06	2.70e-05	2.58e-05	1.21e-05	4.30e-06	2.0e-06	1.1e-06	3.0e-07	1e-07
Marche	M	8.90e-06	3.08e-05	1.68e-05	1.15e-05	4.60e-06	2.4e-06	1.1e-06	4.0e-07	2e-07
Lazio	M	8.90e-06	2.72e-05	2.00e-05	1.15e-05	5.10e-06	2.5e-06	9.0e-07	6.0e-07	2e-07
Abruzzo	S	7.60e-06	2.79e-05	2.29e-05	1.11e-05	3.40e-06	2.3e-06	3.0e-07	8.0e-07	4e-07
Molise	S	9.50e-06	2.74e-05	2.13e-05	8.40e-06	6.20e-06	1.7e-06	1.7e-06	4.0e-07	3e-07
Campania	S	1.83e-05	3.71e-05	1.29e-05	5.10e-06	1.50e-06	1.2e-06	5.0e-07	1.0e-07	1e-07
Puglia	S	1.82e-05	3.39e-05	1.51e-05	5.60e-06	1.90e-06	5.0e-07	1.0e-06	3.0e-07	2e-07
Basilicata	S	1.89e-05	2.96e-05	1.30e-05	8.90e-06	3.00e-06	2.4e-06	5.0e-07	4.0e-07	2e-07
Calabria	S	1.77e-05	3.21e-05	1.61e-05	6.00e-06	3.20e-06	4.0e-07	8.0e-07	2.0e-07	2e-07
Sicilia	S	1.90e-05	3.66e-05	1.26e-05	4.10e-06	2.30e-06	1.1e-06	9.0e-07	2.0e-07	2e-07
Sardegna	S	1.28e-05	3.26e-05	1.66e-05	9.40e-06	3.40e-06	1.1e-06	4.0e-07	2.0e-07	1e-07
Midpoints of intervals t_{ij}		6574	19591	32608	45625	58641	71658	84675	97692	110709

Table 6: Histogram data in $N = 20$ Italian regions: $\mathbf{y}_i$ are raw density values at interval midpoints $\mathbf{t}_i = (t_{i1}, \ldots, t_{iD})'$ of $D = 9$ income classes.

Region	loc.	Raw clr density values, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iD})', i = 1, \ldots, N$								
Piemonte	N	0.534	2.339	2.167	1.276	0.336	-0.545	-1.432	-1.992	-2.685
Valle d'Aosta	N	-0.780	2.241	2.059	1.654	0.044	-1.026	-1.369	-1.052	-1.771
Lombardia	N	0.273	1.399	1.390	0.795	0.460	-0.167	-1.127	-1.309	-1.715
Trentino	N	-0.011	1.799	1.683	0.882	0.927	-0.599	-0.263	-2.209	-2.209
Veneto	N	0.627	1.845	1.590	1.220	0.438	-0.794	-1.236	-1.642	-2.047
Friuli	N	0.401	1.739	1.417	1.095	0.448	0.190	-0.697	-1.950	-2.643
Liguria	N	0.643	2.156	1.669	1.669	0.683	-0.456	-0.456	-2.646	-3.263
Emilia Romagna	N	0.099	1.686	1.598	1.213	0.409	-0.219	-1.200	-1.287	-2.299
Toscana	M	-0.295	1.616	1.656	1.222	0.645	-0.080	-1.317	-1.317	-2.128
Umbria	M	0.348	2.252	2.209	1.447	0.417	-0.345	-0.905	-2.291	-3.132
Marche	M	0.961	2.206	1.603	1.223	0.317	-0.330	-1.119	-2.035	-2.826
Lazio	M	0.943	2.063	1.756	1.204	0.389	-0.342	-1.298	-1.808	-2.907
Abruzzo	S	0.860	2.154	1.958	1.231	0.061	-0.344	-2.342	-1.443	-2.136
Molise	S	0.991	2.049	1.795	0.865	0.555	-0.744	-0.744	-2.190	-2.577
Campania	S	2.260	2.970	1.916	0.976	-0.253	-0.435	-1.351	-2.738	-3.345
Puglia	S	2.021	2.640	1.833	0.832	-0.235	-1.535	-0.842	-1.940	-2.774
Basilicata	S	1.877	2.323	1.502	1.119	0.020	-0.203	-1.840	-2.048	-2.750
Calabria	S	1.987	2.585	1.892	0.911	0.282	-1.797	-1.104	-2.316	-2.440
Sicilia	S	2.114	2.771	1.708	0.573	-0.014	-0.777	-0.931	-2.722	-2.722
Sardegna	S	1.670	2.603	1.931	1.364	0.335	-0.764	-1.680	-2.529	-2.930
Midpoints of intervals t_{ij}		6574	19591	32608	45625	58641	71658	84675	97692	110709

Table 7: Input data for smoothing procedure: $\mathbf{z}_i$ are raw clr density values at interval midpoints $\mathbf{t}_i = (t_{i1}, \ldots, t_{iD})'$ of $D = 9$ income classes.

Appendix C: Aggregated concentrations of metabolite C18

weight[g]	log(weight)	Proportions of log(C18) classes, $\mathbf{W}_i = (W_{i1}, \dots, W_{iD})', i = 1, \dots, D$									
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 8: Proportions of $D = 10$ log(C18) classes within $N = 10$ log(weight) groups for girls. The values $t_j, j = 1, \dots, D$ are the midpoints of the log(C18) subintervals of $\Omega_b = [-2.936, -0.939]$.

weight[g]	log(weight)	Proportions of log(C18) classes, $\mathbf{W}_i = (W_{i1}, \dots, W_{iD})', i = 1, \dots, D$									
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 9: Proportions of $D = 10$ log(C18) classes within $N = 10$ log(weight) groups for boys. The values $t_j, j = 1, \dots, D$ are midpoints of log(C18) subintervals of $\Omega_b = [-2.813, -0.763]$.

weight[g]	log(weight)	Clr transformation of log(C18) classes, $\mathbf{z}_i = (z_{i1},\dots,z_{iD})', i = 1,\dots,N$									
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 10: Clr transformation of $D = 10$ log(C18) classes within the $N = 10$ log(weight) groups for girls. The values $t_j, j = 1,\dots,D$ are the midpoints of the log(C18) subintervals of $\Omega_g = [-2.936, -0.939]$.

weight[g]	log(weight)	Clr transformation of log(C18) classes, $\mathbf{z}_i = (z_{i1},\dots,z_{iD})', i = 1,\dots,N$									
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 11: Clr transformation of $D = 10$ log(C18) classes within the $N = 10$ log(weight) groups for boys. The values $t_j, j = 1,\dots,D$ are the midpoints of the log(C18) subintervals of $\Omega_b = [-2.813, -0.763]$.

PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISERTATION THESIS SUMMARY

Bayes spaces and their applications

Supervisor: **doc. RNDr. Karel Hron, Ph.D.**
Author: **Mgr. Renáta Talská**

Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: Full-time
The year of submission: 2020

The dissertation thesis was carried out under the full-time postgradual programme Applied Mathematics, field Applied Mathematics, in the Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc.

Applicant: **Mgr. Renáta Talská**
Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University Olomouc

Supervisor: **Doc. RNDr. Karel Hron, Ph.D.**
Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University Olomouc

Reviewers: **Univ.-Prof. Dipl.-Ing. Peter Filzmoser, Dr.techn.**
Institute of Statistics and Mathematical Methods in Economics
Faculty of Mathematics and Geoinformation
Vienna University of Technology

Prof. RNDr. Jaromír Antoch, CSc.
Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University Prague

Dissertation thesis summary was sent to distribution on ....

Oral defence of dissertation thesis will be performed on .... at Department of Mathematical Analysis and Applications of Mathematics in front of the committee for Ph.D. study programme Applied Mathematics, Faculty of Science, Palacký University Olomouc, room ......, 17. listopadu 12, Olomouc. Full text of the dissertation thesis is available at Study Department of Faculty of Science, Palacký University Olomouc.

Full text of the doctoral thesis is available at Study Department of Faculty of Science, Palacký University Olomouc.

Contents

1 Abstract

Probability density functions (PDFs) are understood as functional data carrying relative information. Their features such as scale invariance and relative scale are well captured by the theory of Bayes spaces of measures; Bayes spaces thus represent a generalization of the Aitchison geometry for compositional data. These spaces have the Hilbert space structure whose origin is determined by a given reference measure and it can be easily changed through the well-known chain rule. The algebraic-geometric structure of these spaces enables to express PDFs as real functions in the standard L^2 space with the same reference measure using the centered logratio (clr) transformation. This is key to propose statistical methods for PDFs by adapting popular methods of functional data analysis (FDA) which are typically designed in the L^2 space. Since the resulting transformed PFDs have the zero integral (with respect to the given reference measure), they are elements of a subspace of the L^2 space, hereafter denoted as L_0^2. The thesis aims to introduce Bayes spaces of PDFs on a bounded domain in case of (i) the Lebesgue measure and (ii) a general probability measure, and their application to selected problems of FDA. Similar as in FDA, a proper statistical preprocessing of discretely sampled PDFs is crucial for any further analysis. A novel methodology based on principles of Bayes spaces was developed and to use (smoothing) spline functions called compositional (smoothing) splines. Their construction relies on building up a B-spline basis system directly in the L_0^2 space w.r.t. the Lebesgue reference measure. Consequently, the compositional splines can be implemented into FDA methods for statistical processing of PDFs, as demonstrated in detail in case of regression analysis with functional response formed by PDFs. The thesis further deals with weighting of PDFs through the reference measure. The impact on statistical analysis is illustrated through an application to the functional principal component analysis of Italian income data. For its implementation, as well as for the other methods of FDA, it is essential to use a novel centered logratio transformation that maps Bayes spaces with a general reference measure into the L_0^2 space with the Lebesgue reference measure.

Key words: Bayes spaces, probability density functions, reference measure, centered logratio transformation, compositional splines, B-spline representation, functional regression analysis, functional principal component analysis

2 Abstrakt v českém jazyce

Hustotami rozdělení pravděpodobností (angl. probability density functions, PDFs) rozumíme funkcionální data nesoucí relativní informaci. Jejich vlastnosti jako invariantnost na změnu měřítka a relativní měřítko jsou zachyceny pomocí Bayesových prostorů měr; Bayesovy prostory tak představují zobecnění Aitchisonovy geometrie pro kompoziční data. Tyto prostory mají strukturu Hilbertova prostoru, jehož počátek je dán referenční mírou, která může být jednoduše změněna pomocí známého řetězového pravidla. Algebraická struktura Bayesových prostorů umožňuje PDFs vyjádřit jako reálné funkce ve standardním L^2 prostoru vzhledem ke zvolené referenční míře použitím centrované logpodílové (clr) transformace. Toto je klíčové pro možnost užití metod funkcionální analýzy dat (functional data analysis, FDA) pro statistické zpracování hustot, neboť tyto metody jsou typicky navržené právě v prostorech L^2. Protože výsledné transformované PDFs mají nulový integrál (vzhledem k dané referenční míře), jedná se o prvky podprostoru L^2, který je dále označen jako L_0^2. Cílem této disertační práce je představit Bayesovy prostory jako prostory hustot na omezeném intervalu s (i) Lebesgueovou a (ii) obecnou pravděpodobnostní referenční mírou, a jejich aplikace pro vybrané metody FDA. Podobně jako v FDA, vhodné statistické předzpracování diskrétně pozorovaných PDFs je klíčové pro jejich následnou analýzu. Nová metodika založená na principech Bayesových prostorů navrhuje užití (vyhlazovacích) splajnů nazvaných kompoziční (vyhlazující) splajny. Jejich konstrukce je založena na vytvoření B-splajnového bázového systému přímo v prostoru L_0^2 vzhledem k Lebesgueově referenční míře. Následně mohou být kompoziční splajny implementovány do FDA metod pro statistické zpracování PDFs, což je podrobně demonstrováno na případu regresní analýzy se závisle proměnnou reprezentovanou PDFs. Disertační práce se věnuje i aspektu vážení oboru hodnot hustot prostřednictvím referenční míry. Vliv změny referenční míry na statistickou analýzu PDFs je demonstrován pomocí funkcionální metody hlavních komponent na souboru dat o příjmech v Itálii. Pro její implementování, stejně tak jako dalších metod FDA, je klíčové použití nové clr transformace, která zobrazí Bayesovy prostory s obecnou referenční mírou do L_0^2 prostorů s Lebesgueovou referenční mírou.

Klíčová slova: Bayesovy prostory, hustoty rozdělení pravděpodobností, referenční míra, centrovaná logpodílová transformace, B-splajnová reprezentace, regresní analýza, funkcionální metoda hlavních komponent

3 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, a PDF is a non-negative Borel measurable function constrained to integrate to a constant, conventionally set to one. Several authors [5, 8, 32, 33] noted that PDFs have a *relative* nature, in the sense that the meaningful information is embedded in the relative contribution of the probability of any (Borel) subset of the domain of the random variable generating the data to the overall probability, i.e. the measure of the whole set (so-called *total*). Changing the value of the total by multiplying the PDF by a positive real constant results in a scaled density conveying the same *relative* information (which is known as the *scale invariance* property). As a consequence, the actual total is in fact irrelevant for the purpose of the analysis, as widely recognized in Bayesian statistics [14]. The total used simply determines a representative of the equivalence class of proportional density functions.

The relative nature of PDFs can be explained directly with an example: the relative increase of a probability over a Borel set from 0.05 to 0.1 (2 multiple) differs from the increase 0.5 to 0.55 (1.1 multiple), although the absolute differences are the same in both cases. This is known as the *relative scale* property of PDFs. It motivates the use of the so-called logratio approach – a well-established methodology for the analysis of compositional data. These are vectors describing quantitatively the parts of some whole, and are frequently represented as constrained data (e.g. proportions, percentages) carrying relative information [1, 24]. PDFs can be then interpreted as the continuous counterparts of compositions, i.e., as compositions with infinitely many parts. This has recently motivated the construction of the so-called Bayes Hilbert spaces, whose geometry results from the generalization of the Aitchison geometry for compositional data [10] to the infinite-dimensional case. While the pioneering work on Bayes spaces [8] covers only the case assuming that densities are defined on a finite support, Van den Boogaard et. al [33] extended this concept even for densities on possibly unbounded support and introduced Bayes spaces in a more general setting, i.e. as spaces of measures endowed with the Hilbert space structure. In general, Bayes spaces can be defined only if a reference measure P has been set. In the pioneering work, the reference measure was set by default to the Lebesgue (i.e. uniform) reference measure, however, to deal with unbounded support, a non-uniform reference measure P has to be considered as it shown in the latter work. Although Bayes spaces allow to deal with both unbounded and bounded domains for the PDFs, the latter case has been mainly considered so far in practice, and it will be the main focus in this thesis.

Nowadays, we experience an increasing interest in the development of statistical methods for the analysis of PDFs [2, 15, 16, 20, 22, 23, 25, 26, 28]. Although functional data analysis (FDA) [27] may potentially provide a wide range of methodological tools for this purpose, they are typically designed for data embedded in the L^2 space of square-integrable functions. As such, they can not be applied directly to densities since the metric of L^2 spaces does not account for their peculiar properties (e.g., the aforementioned scale invariance and relative scale). The key point in the analysis of PDFs is to map them from Bayes spaces to L^2 spaces where standard FDA methods (e.g., smoothing of functional data, clustering, regression analysis, functional principal component analysis, etc.) can be applied.

The thesis aims to introduce the concept of Bayes space methodology which turns out to be a relevant approach to statistical analysis of PDFs. Three particular novel approaches to statistical processing of PDFs will be presented such as smoothing of PDFs [18], functional regression with the response variable represented by PDF [31] and weighting in Bayes spaces with implications for dimensionality reduction of PDFs using simplicial functional principal component analysis [30].

4 Recent state summary

4.1 Bayes spaces

Bayes spaces represent an algebraic-geometric structure of equivalence classes of proportional σ-finite measures, including probability measures. An arbitrary σ-finite measure P can be selected as the origin of the space. Once such a measure is stated, all measures can be identified with density functions with respect to the measure P, resulting from considering densities as Radon-Nikodym derivatives. Accordingly, P is referred to as the reference measure. In the following, we restrict our attention only to positive (probability) measures on a bounded domain $\Omega = [a, b] \subset \mathbb{R}$. In this setting, the reference measure is set by default to a uniform measure, i.e. to the Lebesgue measure. In the thesis, also discrete (probability) measures are marginally mentioned, nevertheless, they are skipped in this summary.

Bayes spaces: sample space of scale invariant positive measures

Let's assume that the distribution of a continuous random variable is characterized by σ-finite positive measure μ on measurable space $(\Omega, \mathcal{A})$ with reference measure P, $\Omega = [a, b] \subset \mathbb{R}$ and $\mathcal{A}$ being Borel σ-algebra $B([a, b])$. In this setting, the reference measure P can be set to the Lebesgue measure λ, restricted here to a bounded support. The reference density is then the reference measure with respect to itself, i.e. $d\lambda/d\lambda = 1$ on Ω. Given two measures μ and ν with λ-densities $f = d\mu/d\lambda$ and $g = d\mu/d\lambda$, we say that two measures (densities) are $\mathcal{B}(\lambda)$-equivalent, denoted by $\nu =_{\mathcal{B}(\lambda)} \mu$ ($f =_{\mathcal{B}(\lambda)} g$), if they are proportional. That is, in terms of measures, if there exists a positive real constant c such that, for any subset $\mathrm{B} \in B([a, b])$, $\mu(\mathrm{B}) = c \cdot \nu(\mathrm{B})$. If $\mu(\Omega) = 1$ (i.e., μ is a probability measure), we single out a particular representative within a $\mathcal{B}(\lambda)$-equivalence class of proportional measures (densities) which provides the same *relative* information. Indeed, this is typically quantified through the (log-)ratios $\mu(\mathrm{B}_1)/\mu(\mathrm{B}_2)$, with B_1, B_2 in $B([a, b])$ (equivalently in terms of densities, i.e, $f(t_1)/f(t_2)$, with t_1, t_2 in $\Omega = [a, b]$), which are clearly invariant within the $\mathcal{B}(\lambda)$-equivalence class (i.e. *scale invariance* is followed). Within the concept of Bayes spaces, the only relevant information embedded into measures (densities) itself is the relative one. This motivated the use of the log-ratio approach, already known from (multivariate) compositional data analysis, to deal with density functions.

For a fixed reference measure $\mathsf{P} = \lambda$, Bayes space $\mathcal{B}^2(\lambda)$ is a space of $\mathcal{B}(\lambda)$-equivalence classes of σ-finite positive measures on $\Omega = [a, b]$ with square-integrable log-density with respect to reference measure λ:

$$\mathcal{B}^2(\lambda) = \left\{ \mu \in \mathcal{B}^2(\lambda) : \int \left| \ln \frac{d\mu}{d\lambda} \right|^2 d\,\lambda < +\infty \right\}, \tag{1}$$

where measures are identified with the corresponding Radon-Nikodym densities; or, equivalently, we can say that $\mathcal{B}^2(\lambda)$ consists of $\mathcal{B}(\lambda)$-equivalence classes of proportional density functions $f = \frac{d\mu}{d\lambda}$ on $\Omega = [a, b]$ whose logarithm is square-integrable w.r.t. λ. We note that $\mathcal{B}(\lambda)$ is a space for measures as well as for densities since in both cases they are elements of this space. Nevertheless, whether $\mathcal{B}(\lambda)$ is interpreted as space of densities or measures should be obvious from the context.

4.1.1 Hilbert structure of Bayes spaces

The basic operations named *perturbation* ($\oplus$) and *powering* ($\odot$) represent addition and multiplication in $\mathcal{B}^2(\lambda)$. Moreover, the first of them can be interpreted as Bayes updating which gave the name to these spaces. The operations are defined as follows,

$$(\mu \oplus \nu)(\mathrm{B}) =_{\mathcal{B}(\lambda)} \int_{\mathrm{B}} \frac{d\mu}{d\lambda} \cdot \frac{d\nu}{d\lambda}\, d\lambda, \quad \mathrm{B} \in B, \tag{2}$$

and

$$(\alpha \odot \mu)(\mathrm{B}) =_{\mathcal{B}(\lambda)} \int_{\mathrm{B}} \left(\frac{d\mu}{d\lambda} \right)^{\alpha} d\lambda, \quad \mathrm{B} \in B; \tag{3}$$

where μ and ν are measures in $\mathcal{B}^2(\lambda)$ and α is a real number. The operations (2) and (3) can be equivalently expressed using densities. That is, for $f = \frac{d\mu}{d\lambda}$ and $g = \frac{d\nu}{d\lambda}$ we have that

$$(f \oplus g)(t) =_{\mathcal{B}(\lambda)} f(t) \cdot g(t) \quad \text{and} \quad (\alpha \odot f)(t) =_{\mathcal{B}(\lambda)} f(t)^{\alpha}, \quad t \in \Omega. \tag{4}$$

In [8], it is proven that $\mathcal{B}^2(\lambda)$ equipped with the operations $(\oplus, \odot)$ is a vector space. Note that the neutral elements of perturbation and powering are $e(t) = \frac{1}{\lambda(\Omega)} = \frac{1}{b-a}$ (i.e., the uniform density), and 1, respectively. The operation subtraction $(\ominus)$ between two densities f, g is obtained as perturbation of f with reciprocal of g,

$$(f \ominus g)(t) =_{\mathcal{B}(\lambda)} f(t) \oplus [(-1) \odot g(t)] \quad t \in \Omega. \tag{5}$$

This operation is identified as the Radon–Nikodym derivative of μ with respect to ν, that is $\frac{d\mu}{d\lambda} \cdot \left(\frac{d\nu}{d\lambda}\right)^{-1} = \frac{d\mu}{d\nu}$.

Finally, to endow $\mathcal{B}^2(\lambda)$ with the Hilbert space structure, an inner product is required. Egozcue et al. [8] defined it for the Lebesgue reference measure on $\Omega = [a, b]$ and van den Boogaard et al. [33] extended the definition to any finite reference measure. Accordingly, the Bayes inner product on $\mathcal{B}^2(\lambda)$ is can be defined [30] as

$$\langle f, g \rangle_{\mathcal{B}(\lambda)} = \frac{1}{2\lambda(\Omega)} \int_\Omega \int_\Omega \ln \frac{f(t)}{f(u)} \ln \frac{g(t)}{g(u)} \, d\lambda(t) d\lambda(u), \quad t, u \in \Omega, \tag{6}$$

where $\lambda(\Omega) = b - a$, and the corresponding norm and distance as

$$\|f\|_{\mathcal{B}(\lambda)} = \sqrt{\langle f, f \rangle_{\mathcal{B}(\lambda)}} \quad \text{and} \quad d_{\mathcal{B}(\lambda)}(f, g) = \|f \ominus g\|_{\mathcal{B}(\lambda)} \,. \tag{7}$$

4.2 First steps for a statistical analysis in Bayes spaces

As a first step of any data analysis, one needs to think about the sample space for data embedding. Hilbert spaces are mostly employed for functional data due to their geometric structure which allows an easier extension of multivariate methods. Statistical methods provided by the FDA are mostly developed under the assumption that the data belongs to the Hilbert space of squared-integrable functions with the reference measure defaulty set to the Lebesgue measure λ, denoted as $L^2(\lambda)$. Although continuous density functions can be viewed as functional data, they are characterized by specific features which are not honored by FDA. Nevertheless, as long as the data are embedded in a separable Hilbert space

[8], an isometric mapping can be found which enables to express elements of Bayes spaces as real functions of the $L^2(\lambda)$ space. Subsequently, standard statistical analysis can be performed via FDA while accounting for the Bayes space geometry. Such mapping can be provided by centered logratio (clr) transformation.

4.2.1 Centered logratio transformation

For $\mathsf{P} = \lambda$, the clr mapping represents an isometric isomorphism (i.e. a bijective map preserving distances) between $\mathcal{B}^2(\lambda)$ and $L^2(\lambda)$ spaces and it is defined in [33] for $f \in \mathcal{B}^2(\lambda)$ as

$$f^c(t) = \mathrm{clr}_\lambda(f)(t) = \ln f(t) - \frac{1}{\lambda(\Omega)} \int_\Omega \ln f(u)\, d\lambda(u), \quad t \in \Omega. \tag{8}$$

Apparently, the clr representation allows to use the ordinary geometry of $L^2(\lambda)$ to conduct operations of perturbation (2), powering (3) and inner product (6) for the elements of $\mathcal{B}^2(\lambda)$, while accounting for the specific features captured by the Bayes space. Indeed,

$$\mathrm{clr}_\lambda(f \oplus g) = \mathrm{clr}_\lambda(f)(t) + \mathrm{clr}_\lambda(g)(t), \quad \mathrm{clr}_\lambda(\alpha \odot f)(t) = \alpha \cdot \mathrm{clr}_\lambda(f)(t) \tag{9}$$

and

$$\langle f, g \rangle_{\mathcal{B}^2(\lambda)} = \langle \mathrm{clr}_\lambda(f), \mathrm{clr}_\lambda(g) \rangle_{L^2(\lambda)} = \int_\Omega \mathrm{clr}_\lambda(f)(t) \cdot \mathrm{clr}_\lambda(g)(t) d\lambda(t), \tag{10}$$

where $\langle \cdot, \cdot \rangle_{L^2(\lambda)}$ denotes the inner product in $L^2(\lambda)$. However, due to the construction, the clr transformed densities are characterized by zero-integral constraint (w.r.t. λ),

$$\int_\Omega \mathrm{clr}_\lambda(f)(t)\, d\lambda(t) = \int_\Omega \ln f(t)\, d\lambda(t) - \int_\Omega \frac{1}{\lambda(\Omega)} \int_\Omega \ln f(u)\, d\lambda(u)\, d\lambda(t) = 0, \tag{11}$$

which needs to be taken into account when analyzing clr transformed densities. As the clr space is clearly a subspace of $L^2(\lambda)$, hereafter it is denoted as $L^2_0(\lambda)$. Note that clr transformation represents one-to-one mapping, so it is possible to map densities in $L^2_0(\lambda)$ back to $\mathcal{B}^2(\lambda)$ by using exponential transformation, i.e. $\exp\left[f^c\right](t) = \exp\left[\mathrm{clr}_\lambda(f)\right](t)$. The resulting back-transformed density f can be closed to the unit integral due to the scale invariance feature.

4.2.2 Smoothing of density functions (Approach I)

The second step in FDA is related to the estimation of the underlying N functions $f_1, \ldots, f_N$ from discretized data $(t_{ij}, y_{ij}), i = 1, \ldots, N, j = 1, \ldots, n_i$, where y_{ij} is observation of f_i at t_{ij}. Spline functions are extensively used in FDA for an approximation of nonperiodical functions as they are flexible enough to cover a wide range of their specific behavior, hence they are also a natural choice for density functions. Moreover, using splines for the representation of density functions turned out to be the most appropriate approximative tool as the associated basis coefficients can be directly used for further statistical analysis, i.e., for instance in functional regression with response formed by PDFs.

In case of density functions, they are discretely sampled in terms of histogram data. That is, for each density $f_i(t), t \in \Omega, i = 1, \ldots, N$, one usually observes a positive real vector $\mathbf{W}_i = (W_{i1}, \ldots, W_{iD_i})'$, whose components correspond to the (absolute or relative) frequencies of D_i classes in which the interval Ω is partitioned. Accordingly, the raw density data y_{ij} correspond to interval midpoints t_{ij} of D_i classes obtained by dividing (not necessary normalized) components of $\mathbf{W}_i$ by the length of the respective classes. Note that data $\mathbf{y}_i = (y_{i1}, \ldots, y_{iD_i})', i = 1, \ldots, N$ can be interpreted as discretized density functions, that is, as compositions. Since it is convenient to perform preprocessing of density functions in clr space L_0^2, discrete version of clr transformation is employed to express compositional vectors $\mathbf{y}_i, i = 1, \ldots, N$ in a standard Euclidean space; this yields clr transformed data denoted as $\mathbf{z}_i = (z_{i1}, \ldots, z_{iD_i})', i = 1, \ldots, N$. Consequently, the aim is to estimate (approximate) the underlying continuous clr density functions $\mathrm{clr}_\lambda(f_i), i = 1, \ldots, N$ from raw given data $(t_{ij}, z_{ij}), i = 1, \ldots, N, j = 1, \ldots, D_i$.

A first attempt of constructing a spline representation adapted for clr transformed density functions was proposed in [17] and will be recalled in this section. In the following, it is assumed that a single density function $\mathrm{clr}_\lambda(f)$ is being approximated.

To set the notation, call values

$$\Delta\lambda := \{\lambda_0 = a < \lambda_1 < \ldots < \lambda_g < b = \lambda_{g+1}\} \tag{12}$$

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 $\Omega = [a, b]$ given the knots $\Delta\lambda$. It is known that $\dim(\mathcal{S}_k^{\Delta\lambda}[a, b]) = g+k+1$. For the construction of all basis functions of $\mathcal{S}_k^{\Delta\lambda}[a, b]$, it is necessary to consider some additional knots. Without loss of generality, we here assume that those additional knots are at the boundary, i.e.,

$$\lambda_{-k} = \cdots = \lambda_{-1} = \lambda_0, \quad \lambda_{g+1} = \lambda_{g+2} = \cdots = \lambda_{g+k+1}. \tag{13}$$

Then every spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$ in the L^2 space has a unique representation as (see [3], [4] for details)

$$s_k\left(t\right) = \sum_{i=-k}^{g} b_i B_i^{k+1}\left(t\right), \tag{14}$$

where the vector $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ is the vector of B-spline basis coefficients of $s_k\left(t\right)$ and functions $B_i^{k+1}(t)$, $i = -k, \ldots, g$ are B-spline functions of the same degree k as spline function $s_k\left(t\right)$ forming basis in $\mathcal{S}_k^{\Delta\lambda}[a, b]$. They are defined for $k = 0$ (order 1) by

$$B_i^1(t) = \begin{cases} 1 & \text{if } t \in [\lambda_i, \lambda_{i+1}) \\ 0 & \text{otherwise.} \end{cases}$$

and for k, $k \in \mathbb{N}$, $k \geq 1$, (order $k+1$) by

$$B_i^{k+1}(t) = \frac{t - \lambda_i}{\lambda_{i+k} - \lambda_i} B_i^k(t) + \frac{\lambda_{i+k+1} - t}{\lambda_{i+k+1} - \lambda_{i+1}} B_{i+1}^k(t).$$

In [17], the optimal smoothing problem was represented as a trade-off between smoothing and the least squares approximation. Assume that data (t_j, z_j), $a \leq t_j \leq b$, the weights $w_j^s \geq 0$, $j = 1, \ldots, D$, $D \geq g+1$ and the parameter $\alpha \in (0, 1)$ are given. For an arbitrary $l \in \{1, \ldots, k-1\}$ our aim is to find a smoothing spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$, which minimizes the functional

$$J_l(s_k) = \alpha \sum_{j=1}^{D} w_j \left[z_j - s_k(t_j)\right]^2 + (1-\alpha) \int_a^b \left[s_k^{(l)}(t)\right]^2 \, dt, \tag{15}$$

and fulfills the condition

$$\int\limits_a^b s_k(t) \, dt = 0, \tag{16}$$

resulting from the clr transformation. The minimization problem (15) represents a compromise between staying close to the given data and obtaining a smooth function. The smoothness of the resulting approximation is affected by the smoothing parameters α and l, where l stands for lth derivative. Similarly, one can minimize the following functional with respect to condition (16) for some positive parameter α, i.e.

$$J_l(s_k) = \sum_{j=1}^{D} w_j \left[z_j - s_k(t_j)\right]^2 + \alpha \int_a^b \left[s_k^{(l)}(t)\right]^2 \, dt \tag{17}$$

For the sake of brevity, we will focus on the minimization of the functional (17). It was proven that the *optimal* smoothing spline for this task is the spline $s_k^*(t)$, given by formula

$$s_k^*\left(t\right) = \sum_{i=-k}^{g} b_i^* B_i^{k+1}\left(t\right), \tag{18}$$

with B-spline coefficients $\mathbf{b}^* = (b_{-k}^*, \ldots, b_g^*)'$ obtained as (see [17] for details)

$$\mathbf{b}^* = \mathbf{V}\mathbf{z} \tag{19}$$

with

$$\mathbf{V} := \mathbf{DK}\left[\alpha\left(\mathbf{DK}\right)'\mathbf{N}_{kl}\mathbf{DK} + \left(\mathbf{B}_{k+1}(\mathbf{t})\mathbf{DK}\right)'\mathbf{W}^s\mathbf{B}_{k+1}(\mathbf{t})\mathbf{DK}\right]^+\mathbf{K}'\mathbf{D}\mathbf{B}_{k+1}'(\mathbf{x})\mathbf{W}^s. \tag{20}$$

Here $\mathbf{A}^+$ denotes the Moore-Penrose pseudoinverse of a matrix $\mathbf{A}$, $\mathbf{W}^s = \mathrm{diag}(\mathbf{w}^s)$, $\mathbf{w}^s = (w_1^s, \ldots, w_D^s)'$, $\mathbf{t} = (t_1, \ldots, t_D)'$, $\mathbf{z} = (z_1, \ldots, z_D)'$,

$$\mathbf{B}_{k+1}(\mathbf{t}) = \begin{pmatrix} B_{-k}^{k+1}\left(t_1\right) & \ldots & B_g^{k+1}\left(t_1\right) \\ \vdots & \ddots & \vdots \\ B_{-k}^{k+1}\left(t_D\right) & \ldots & B_g^{k+1}\left(t_D\right) \end{pmatrix} \in \mathbb{R}^{D,g+k+1} \tag{21}$$

is the collocation matrix,

$$\mathbf{D} = (k+1)\,\mathrm{diag}\left(\frac{1}{\lambda_1 - \lambda_{-k}}, \ldots, \frac{1}{\lambda_{g+k+1} - \lambda_g}\right) \in \mathbb{R}^{g+k+1,g+k+1} \tag{22}$$

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

$$\mathbf{M}_{kl} = \begin{pmatrix} \left\langle B_{-k+l}^{k+1-l}, B_{-k+l}^{k+1-l} \right\rangle_{L^2(\lambda)} & \cdots & \left\langle B_{g}^{k+1-l}, B_{-k+l}^{k+1-l} \right\rangle_{L^2(\lambda)} \\ \vdots & & \vdots \\ \left\langle B_{-k+l}^{k+1-l}, B_{g}^{k+1-l} \right\rangle_{L^2(\lambda)} & \cdots & \left\langle B_{g}^{k+1-l}, B_{g}^{k+1-l} \right\rangle_{L^2(\lambda)} \end{pmatrix} \in \mathbb{R}^{g+k+1-l, g+k+1-l}, \tag{23}$$

where

$$\left\langle B_i^{k+1-l}, B_j^{k+1-l} \right\rangle_{L^2(\lambda)} = \int_a^b B_i^{k+1-l}(t) B_j^{k+1-l}(t)\, dt$$

stands for scalar product of B-splines in $L^2(\lambda)$ space. The matrix $\mathbf{S}_l$ is an upper triangular matrix such that

$$\mathbf{S}_l = \mathbf{D}_l\mathbf{L}_l \dots \mathbf{D}_1\mathbf{L}_1 \in \mathbb{R}^{g+k+1-l, g+k+1}, \tag{24}$$

and $\mathbf{D}_{j'} \in \mathbb{R}^{g+k+1-j, g+k+1-j'}$ is a diagonal matrix such that

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

with

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

and

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

5 Thesis objectives

The thesis aims to introduce the concept of Bayes space methodology which turns out to be a relevant approach to statistical analysis of PDFs. Three particular novel approaches to statistical processing of PDFs will be presented such as smoothing of PDFs [18], functional regression with the response variable represented by PDF [31] and weighting in Bayes spaces with implications for dimensionality reduction of PDFs using simplicial functional principal component analysis [30]. The potential of the methodological developments is shown on simulated data and real-world data (illustrative examples are skipped in this summary).

6 Theoretical framework and applied methods

6.1 Smoothing of density functions (Approach II)

We note that an initial approach for approximation of density functions in clr space has some limitations since the basis elements do not belong to the L_0^2 space. Therefore, an important step ahead is made by constructing a B-spline basis directly in the clr space L_0^2. As a direct consequence, the B-splines can be expressed directly in Bayes spaces leading to spline representation of density functions in the original space; hereafter we refer to *compositional splines*. Apart from methodological advantages, using compositional splines simplifies the construction and interpretation of spline coefficients that can be considered as coefficients of a basis in Bayes spaces.

Let the sequence of knots (12) is given. We define the functions $Z_i^{k+1}(t)$ for $k \geq 0$, $k \in \mathbb{N}$, which are the first derivatives of the B-splines $B_i^{k+2}(t)$ for $k \geq 0$, $k \in \mathbb{N}$, as

$$Z_i^{k+1}(t) := \frac{d}{dt} B_i^{k+2}(t), \tag{25}$$

i.e., more precisely for $k = 0$

$$Z_i^1(x) = \begin{cases} 1 & \text{if } x \in [\lambda_i, \lambda_{i+1}) \\ -1 & \text{if } x \in (\lambda_{i+1}, \lambda_{i+2}] \end{cases}$$

and for $k \geq 1$

$$Z_i^{k+1}(t) = (k+1)\left(\frac{B_i^{k+1}(t)}{\lambda_{i+k+1} - \lambda_i} - \frac{B_{i+1}^{k+1}(t)}{\lambda_{i+k+2} - \lambda_{i+1}}\right). \tag{26}$$

The functions $Z_i^{k+1}(t)$ have similar properties as B-spline functions $B_i^{k+1}(t)$. They are piecewise polynomials of degree k on local support for $k \geq 1$,

$$\text{supp } Z_i^{k+1}(t) \ = \ \text{supp } B_i^{k+2}(t) = [\lambda_i, \lambda_{i+k+2}],$$

with continuous derivatives up to order $k-1$. From the perspective of L_0^2 space, a crucial point is that the integral of $Z_i^{k+1}(t)$ equals to zero. If we consider Curry-

Schoenberg B-spline $M_i^{k+1}(t)$ [3], which are defined as

$$M_i^{k+1}(t) \;:=\; \frac{k+1}{\lambda_{i+k+1}-\lambda_i}\, B_i^{k+1}(t)$$

with property

$$\int\limits_{\mathbb{R}} M_i^{k+1}(t)\, dt \;=\; 1,$$

than it is clear that

$$Z_i^{k+1}(t) \;=\; M_i^{k+1}(x) - M_{i+1}^{k+1}(t) \tag{27}$$

and

$$\int\limits_{\mathbb{R}} Z_i^{k+1}(t)\, dt \;=\; 0.$$

Now, regarding the definition (25), we are able to use spline functions $Z_i^{k+1}(t)$ which have zero integral on Ω (denoted as ZB-splines in the sequel). In the following, $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ denotes the vector space of polynomial splines of degree $k > 0$, defined on a finite interval $\Omega = [a,b]$ with the sequence of knots $\Delta\lambda$ given in (12) and having zero integral on $[a,b]$, it means

$$\mathcal{Z}_k^{\Delta\lambda}[a,b] \;:=\; \left\{ s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b] \;:\; \int\limits_I s_k(t)\, dt = 0 \right\}. \tag{28}$$

Theorem 6.1 *The dimension of the vector space $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ defined by the formula (28) is $g+k$.*

Theorem 6.2 *For the coincident additional knots (13), the functions $Z_{-k}^{k+1}(t)$, $\cdots$, $Z_{g-1}^{k+1}(t)$ form a basis for the space $\mathcal{Z}_k^{\Delta\lambda}[a,b]$.*

With regard to this theorem, each spline $s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}[a,b]$ has a unique representation

$$s_k\,(t) = \sum_{i=-k}^{g-1} b_i^z Z_i^{k+1}\,(t)\,. \tag{29}$$

Now we can proceed to a matrix notation of $s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}[a,b]$. With respect to (26) and (27), we are able to write the functions $Z_i^{k+1}(t)$ in matrix notation as

$$Z_i^{k+1}(t) = (k+1)\left(B_i^{k+1}(t), B_{i+1}^{k+1}(t)\right) \begin{pmatrix} \dfrac{1}{\lambda_{i+k+1}-\lambda_i} & 0 \\ 0 & \dfrac{1}{\lambda_{i+k+2}-\lambda_{i+1}} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

that is, for ZB-spline basis of $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ we have

$$\left(Z_{-k}^{k+1}(t), \ldots, Z_{g-1}^{k+1}(t)\right) = \left(B_{-k}^{k+1}(t), \ldots, B_g^{k+1}(t)\right)\mathbf{D}\mathbf{K}^z = \mathbf{B}_{k+1}(t)\mathbf{D}\mathbf{K}^z,$$

where matrix $\mathbf{D}$ is given in (22) and

$$\mathbf{K}^z = \begin{pmatrix} 1 & 0 & 0 \cdots & 0 & 0 \\ -1 & 1 & 0 \cdots & 0 & 0 \\ 0 & -1 & 1 \cdots & 0 & 0 \\ \vdots & \vdots & \ddots \quad \ddots & \vdots & \vdots \\ 0 & 0 & 0 \cdots & -1 & 1 \\ 0 & 0 & 0 \cdots & 0 & -1 \end{pmatrix} \in \mathbb{R}^{g+k+1,g+k}. \tag{30}$$

It follows that each spline $s_k(t)$ from $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ can be written in the matrix notation using the standard B-spline basis with B-spline coefficients $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ fulfilling the condition (45) from Theorem 6.3 as

$$s_k(t) \;=\; \mathbf{Z}_{k+1}(t)\mathbf{b}^z \;=\; \mathbf{B}_{k+1}(t)\mathbf{D}\mathbf{K}^z\mathbf{b}^z = \mathbf{B}_{k+1}(t)\mathbf{b}, \tag{31}$$

with $\mathbf{Z}_{k+1}(t) = \left(Z_{-k}^{k+1}(t), \ldots, Z_{g-1}^{k+1}(t)\right)$ and $\mathbf{b}^z = \left(b_{-k}^z, \ldots, b_{g-1}^z\right)'$. Note that the formula (31) provides a guideline how to convert the splines from $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ to splines with zero integral (with coefficients fulfilling (45)) from $\mathcal{S}_k^{\Delta\lambda}[a,b]$. This is particularly useful from the practical point of view as it allows to use existing codes in the statistical softwares for actual computations of the methods of FDA. For instance, the package `fda` of the statistical software R implements functional principal component analysis for a standard B-spline basis representation of functional data, hence it can be used for sampled (clr) density functions as in Section 5.

Unlike Approach I where the smoothing spline functions with zero integral were constructed upon standard B-spline basis system of functions $B_i^{k+1}(t)$,

$i = -k, \dots, g$, we can now use ZB-spline basis system of functions $Z_i^{k+1}(t)$, $i = -k, \dots, g-1$ for this purpose. Accordingly, for an arbitrary $l \in \{1, \dots, k-1\}$ we aim to find a smoothing spline $s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}[a,b] \subset L_0^2([a,b])$ which minimizes the functional (15). The minimum is found for the spline $s_k^*(t)$ of the form (29) with ZB-spline coefficients $\mathbf{b}^{z*} = \left(b_{-k}^{z*}, \dots, b_{g-1}^{z*}\right)'$ obtained as (see [18] for details)

$$\mathbf{b}^{z*} = \mathbf{V}^z \mathbf{z} \tag{32}$$

with

$$\mathbf{V}^z := \mathbf{G}^{-1}\mathbf{g}, \tag{33}$$

where

$$\mathbf{G} \;:=\; (\mathbf{K}^z)'\mathbf{D}\left[(1-\alpha)\mathbf{S}_l'\mathbf{M}_{kl}\mathbf{S}_l + \alpha\mathbf{B}_{k+1}'(\mathbf{t})\mathbf{W}^{\mathbf{s}}\mathbf{B}_{k+1}(\mathbf{t})\right]\mathbf{D}\mathbf{K}^z \tag{34}$$

and

$$\mathbf{g} \;:=\; \alpha(\mathbf{K}^z)'\mathbf{D}\mathbf{B}_{k+1}'(\mathbf{t})\mathbf{W}^s;$$

the matrices $\mathbf{B}_{k+1}(\mathbf{t}), \mathbf{D}, \mathbf{M}_{kl}, \mathbf{S}_l, \mathbf{K}^z$ are given in (21), (22), (23), (24), (30). Consequently, by considering the formula (31), the resulting smoothing spline in matrix notation using standard B-splines $B_i^{k+1}(x)$ is obtained as

$$s_k^*(t) \;=\; \mathbf{B}_{k+1}(t)\mathbf{D}\mathbf{K}^z\mathbf{b}^{z*},$$

where the vector $\mathbf{b}^{z*}$ is given in (32).

In some applications, the orthonormalization of the B-spline basis might be useful. Note that ZB-spline functions forming the basis system of $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ are by the default setting (26) non-orthogonal. The orthogonalized ZB-spline functions is discussed in the thesis and it is not shown in this summary.

Compositional splines in the Bayes spaces $\mathcal{B}^2(\lambda)$: Construction of spline functions directly in $L_0^2(\lambda)$ has important practical consequences, however, it is crucial also from the theoretical perspective. Expressing B-spline functions as functions in $L_0^2(\lambda)$ enables to transform them back to the original Bayes space $\mathcal{B}^2(\lambda)$ by using the exponential. It results in *compositional B-splines (CB-splines)*, obtained from (26) as

$$\zeta_i^{k+1}(t) =_{\mathcal{B}(\lambda)} \exp[Z_i^{k+1}](t), \quad i = -k, \dots, g-1, \; k \geq 0. \tag{35}$$

As a consequence, it is immediate to define vector space $\mathcal{C}_k^{\Delta\lambda}[a,b]$ of compositional polynomial spline functions of degree $k > 0$, defined on a finite interval $\Omega = [a,b]$ with the sequence of knots $\Delta\lambda$. From isomorphism between $\mathcal{C}_k^{\Delta\lambda}[a,b]$ and $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ it holds that

$$\dim\left(\mathcal{C}_k^{\Delta\lambda}[a,b]\right) = g + k.$$

Moreover, from isometric properties of clr transformation (9) and (10) it follows that each compositional spline function $\xi_k(t) \in \mathcal{C}_k^{\Delta\lambda}[a,b]$ in $\mathcal{B}^2(\lambda)$ can be uniquely represented as

$$\xi_k(t) = \bigoplus_{i=-k}^{g-1} b_i^z \odot \zeta_i^{k+1}(t). \tag{36}$$

The resulting compositional splines (with either orthogonal, or non-orthogonal CB-spline basis system) can be used for the representation of density functions directly in Bayes spaces. This is an important step in the construction of the FDA methods involving density functions. With CB-splines one has a guarantee that methods are developed consistently in the Bayes spaces. Moreover, the possibility of having an orthogonal basis enables to gain additional features resulting from orthogonality of finite dimensional projection in combination with approximate properties of spline functions.

6.2 Statistical methods in Bayes spaces: Functional regression

Regression analysis is a key statistical tool to model a linear relationship between a response variable and a set of covariates. If the response or the predictors have functional nature, the functional regression analysis is to be considered. Although the general problem of functional regression has been extensively studied in the literature on FDA (i.e., for instance in [11, 27, 29]), a concise methodology for regression analysis in the presence of a distributional response has been proposed only recently in [31]. It aims to develop a general theoretical and computational setting allowing for the estimation and uncertainty assessment in linear models with a distributional response. In particular, we focus on a function-on-scalar model for a distributional in Bayes spaces on closed interval $\Omega = [a, b]$. Similarly as in the L^2 setting, the key is to consider the B-spline representation of the PDF response observed as discrete (histogram) data. On these bases, the effective computational procedure is proposed and further discussed in this section.

A function-on-scalar regression model in $\mathcal{B}^2(\lambda)$ is introduced as a counterpart of a model in the L^2 space. We assume the dependent variable $y(t), t \in \Omega$ to be an element of $\mathcal{B}^2(\lambda)$ 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 $\mathcal{B}^2(\lambda)$ of the form

$$y_i(t) = \beta_0(t) \oplus \bigoplus_{j=1}^{r} \left[x_{ij} \odot \beta_j\right](t) \oplus \varepsilon_i(t) \tag{37}$$

where ε_i denotes a zero-mean functional error (or residual) in $\mathcal{B}^2(\lambda)$, $i = 1, \ldots, N$, and the unknown functions β_j, $j = 0, ..., r$, belong to $\mathcal{B}^2(\lambda)$ 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 $\mathcal{B}^2(\lambda)$

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\varepsilon_i\|^2_{\mathcal{B}^2(\lambda)} = \sum_{i=1}^{N} \left\| \bigoplus_{j=0}^{r} \left[x_{ij} \odot \beta_j\right] \ominus y_i \right\|^2_{\mathcal{B}^2(\lambda)}. \tag{38}$$

Note that SSE (38) in the Bayes Hilbert space $\mathcal{B}^2(\lambda)$ represents the analogue of compositional SSE in $\mathcal{B}^2$ space with the discrete uniform reference measure [7] in infinite dimensions. Applying the clr transformation (9) to both sides of the model (37) yields

$$\mathrm{clr}_\lambda(y_i)(t) = \mathrm{clr}_\lambda(\beta_0)(t) + \sum_{j=1}^{r} \left[x_{ij} \cdot \mathrm{clr}_\lambda(\beta_j)\right](t) + \mathrm{clr}_\lambda(\varepsilon_i)(t), \quad i = 1, \ldots, N, \tag{39}$$

that enables one to reformulate the objective SSE (38) equivalently in the L^2 sense as

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\mathrm{clr}_\lambda(\varepsilon_i)\|^2_{L^2(\lambda)} = \sum_{i=1}^{N} \left\| \sum_{j=0}^{r} \left[x_{ij} \cdot \mathrm{clr}_\lambda(\beta_j)\right] - \mathrm{clr}_\lambda(y_i) \right\|^2_{L^2(\lambda)}. \tag{40}$$

In this thesis, the focus is on SSE, not on penalized SSE, since one may control the smoothness of the estimated functions for $\mathrm{clr}_\lambda(\beta_j(t))$ through the smoothness of the B-spline representation of the response, as shall be discussed further in this section.

As a next step, since both the observed functions $\mathrm{clr}_\lambda(y_i)(t)$, $i = 1, \ldots, N$, and regression parameters $\mathrm{clr}_\lambda(\beta_j)(t)$, $j = 0, \ldots, r$, are elements of $L^2_0(\lambda)$, their basis expansion fulfilling the zero-integral constraint on Ω using a given basis system $\{\varphi_k, k = 1, ..., K\}$ must be considered, i.e.,

$$\int_I \mathrm{clr}_\lambda(y_i(t))dt = \int_I \sum_{k=1}^{K} c_{ik}\varphi_k(t)dt = 0; \quad \int_I \mathrm{clr}_\lambda(\beta_j(t))dt = \int_I \sum_{k=1}^{K} b_{jk}\varphi_k(t)dt = 0. \tag{41}$$

Both approaches to basis expansions designed for densities, i.e., Approach I (4.2.2) and Approach II 6.1, respectively, can be used when estimating the linear model (37). Nevertheless, regarding to the key results proved in the following section (Theorem 6.3), the former approach leads to constraints on the coefficients $\{c_{ik}\}$, $\{b_{jk}\}$ and consequently on model singularities. Although this can be overcome by using the latter method based on compositional splines, both of them in fact lead to the same estimations of regression parameters. Therefore, in the following we will mainly focus on the consequences related to the former approach as developed in [31].

6.2.1 Regression modeling of B-spline coefficients using Approach I

Let us consider the B-spline representations for the clr transformed observations of the response density, i.e., $\mathrm{clr}_\lambda(y_i)(t), i = 1, \ldots, N$, of the form

$$s_k^i\left(t\right) = \sum_{j=-k}^{g} Y_{i,j+k+1} B_j^{k+1}\left(t\right), \tag{42}$$

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

$$\mathbf{Y}_{(i)} = \mathbf{V}\mathbf{z}_{(i)}, \quad i = 1, \ldots, N; \tag{43}$$

the matrix $\mathbf{V}$ of dimensions $(g+k+1)\times D$ is given in (20) and $\mathbf{z}_{(i)} = (z_{i1}, \ldots, z_{iD})'$, $i = 1, \ldots, N$ are vectors of clr transformed raw density data. If the same B-spline basis system is used for all the data, (43) can be expressed in matrix notation as

$$\underline{\mathbf{Y}} \;=\; \underline{\mathbf{Z}}\mathbf{V}', \tag{44}$$

where $\underline{\mathbf{Y}}$, $\underline{\mathbf{Z}}$ are the matrices of dimensions $N \times (g+k+1)$ and $N \times D$, respectively, having the following form,

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

The explicit expression for the optimal smoothing B-spline is given in (18). As an element of innovation, we aim to find the necessary and sufficient condition for the vector $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ to be the vector of B-spline coefficients for spline with zero integral. The following Theorem 6.3 characterizes all the splines with zero integral (not necessarily a smoothing spline) using a standard B-spline basis system through a necessary and sufficient condition on the vector $\mathbf{b}$.

Theorem 6.3 *For a spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$, $s_k\left(t\right) = \sum\limits_{i=-k}^{g} b_i B_i^{k+1}\left(t\right)$, the condition $\int\limits_a^b s_k(t)\, dt = 0$ is fulfilled if and only if*

$$\sum_{i=-k}^{g} b_i \left(\lambda_{i+k+1} - \lambda_i\right) \;=\; 0. \tag{45}$$

In the light of Theorem 6.3, it is easy to see that vector $\mathbf{b}$ is orthogonal to the vector $(\lambda_1 - \lambda_{-k}, \ldots, \lambda_{g+k+1} - \lambda_g)'$, which only depends on the knots positions. Further, the vectors $\mathbf{Y}_{(i)}, i = 1 \ldots, N$, of B-spline coefficients, one has the linear constraints

$$\sum_{j=1}^{g+k+1} Y_{ij} \left(\lambda_j - \lambda_{j-k-1}\right) \; = \; 0. \tag{46}$$

Whenever the same B-spline basis is employed for all the data – as it is usually the case – the linear constraint (46) turns into a model singularity, as we shall show further in this subsection.

By considering the B-spline representations of the clr transformed response functions $\mathrm{clr}_\lambda(y_i)(t), i = 1, \ldots, N$, we can express the model (37) in the form of a multivariate regression model. Following the given notation, spline coefficients for the i-th observation $y_i(t)$ are denoted by $\mathbf{Y}_{(i)} = (Y_{i,1}, \ldots, Y_{i,g+k+1})'$, $i = 1, 2, \ldots, N$, and 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 (37) the multivariate linear regression model of the form

$$\underline{\mathbf{Y}}_{(N\times(g+k+1))} = \mathbf{X}_{(N\times p)}\mathbf{B}_{(p\times(g+k+1))} + \underline{\boldsymbol{\varepsilon}}_{(N\times(g+k+1))}, \tag{47}$$

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

$$\widehat{\mathbf{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{g+k+1}), \tag{48}$$

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

$$\widehat{\boldsymbol{\beta}}_j = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y}_j,\ \ j = 1, \ldots, g+k+1. \tag{49}$$

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., $\mathcal{M}(\mathbf{X}) = \{\mathbf{X}\mathbf{u} : \mathbf{u} \in \mathbb{R}^p\}$.

Because the realization of the 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 (see (46)) due to Theorem 6.3. Accordingly, one may expect that a similar constraint applies to the corresponding estimated regression coefficients, as stated by the following result.

Proposition 6.2.1 *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* $s = 0, \ldots, r$*.*

The latter constraint introduces a singularity into the regression model (47), which may affect parameter inference. Similarly as in multivariate regression [13], 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. Since vectors $\mathbf{Y}_{(i)}$, $i = 1, ..., N$, form a hyperplane $\mathcal{H}$ of dimension $g+k$, orthogonal to the normal vector $(\lambda_1 - \lambda_{-k}, \ldots, \lambda_{g+k+1} - \lambda_g)'$ 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 (41) – 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, that in turn correspond to the

Simplicial Functional Principal Components (SFPCs) of the smoothed densities $y_1(t), ..., y_N(t)$ [15]. However, note that the BLUE estimation (48) 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. Of course, the singularity problem can be prevented by considering ZB-spline basis system from $L_0^2(\lambda)$, so that the response is expressed through a set of unconstrained coefficients, namely ZB-spline coefficients.

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 from Section 3.2 is applied (hereafter named "regression-smoothing"), and (b) a compositional regression [7] is applied, estimating the model

$$\mathbf{z}_i \;=\; \boldsymbol{\beta}_0^{(Z)} + \sum_{j=1}^{r} \boldsymbol{\beta}_j^{(Z)} x_{ij} + \boldsymbol{\epsilon}_i, \tag{50}$$

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 presented here and that one proposed in [7]

$$\begin{tikzcd}
\underline{\mathbf{Z}} \arrow[r, "smoothing"] \arrow[d, "regression"'] & \underline{\mathbf{Y}} \arrow[d, "regression"] \\
\underline{\widehat{\mathbf{Z}}} \arrow[r, "smoothing"'] & \underline{\widehat{\mathbf{Y}}}
\end{tikzcd} \tag{51}$$

It can be shown that under particular conditions the "smoothing-regression" and "regression-smoothing" approaches are equivalent. Nevertheless, the proposed framework provides a much 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 to the same set of classes, which may not be the optimal one 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, [21, 22, 23]). Dealing with high-dimensional (compositional) data from a discrete viewpoint may yield issues related to the curse of dimensionality, which are completely overcome with a functional viewpoint.

Moreover, 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 penalized SSE criterion [27].

The thesis also deals with assessing the goodness-of-fit of the model on the observed density curves via coefficient of determination and uncertainty in the estimation of regression parameters is incorporated based on a resampling method (bootstrap). Both are detailed in the thesis.

6.3 Weighted Bayes spaces

Weighted Bayes spaces refer to Bayes spaces with the reference measure other than the uniform one. The name *weighted Bayes spaces* reflects the fact that changing the uniform reference measure induces a (non-uniform) weighting of the domain of PDFs. Accordingly, Bayes spaces with the uniform reference measures are referred from now on to as *unweighted* Bayes spaces. Linking a weighting scheme to a non-uniform reference measure has been already discussed for multivariate compositions in [9].

The weighting of a domain of PDFs can be relevant in practice, as rarely all regions of the domain (compositional parts in the multivariate case) have the same importance or relevance for the analysis. For example, it is known that in particle-size distributions [21, 22, 23], finer fractions of soil are measured for some methods with lower reliability than crude fractions [12], which implies naturally higher relevance of the latter and their respective subdomain. Another example is represented by income distributions across various regions (see Section 5.2). The lower-income values are going to be of primary interest for policy makers when the aim is to reveal regions suffering from poverty. In addition, here the relative scale, which implies a larger impact of changes in small income values, matters and should be highlighted. And yet another reason why weighting can be convenient is to analyze deviations from a common trend in data. All these cases can benefit from a sensible weighting scheme which gives more relevance to certain regions of the domain of the PDF when conducting functional data analysis.

Weighted Bayes spaces: sample space of scale invariant positive measures

Following the notation from Section 1.1, we now assume that the reference measure of a measurable space $(\Omega, \mathcal{A}) = ([a,b], B([a,b]))$ is fixed to a general (probability) measure P. Then given measure μ with its P-density $f = d\mu/d\mathsf{P}$ (i.e., w.r.t. the reference measure P), the probability measure of any event $\mathrm{B} \in B([a,b])$ is

$$\mu(\mathrm{B}) = \int_{\mathrm{B}} f \, d\mathsf{P} = \int_{\mathrm{B}} \frac{d\mu}{d\mathsf{P}} \, d\mathsf{P}.$$

Note that the choice of the reference measure is not scale invariant, because it reflects on the scale of the entire Bayes space. For instance, the Lebesgue measure on a domain $\Omega = [a,b]$ is proportional to the uniform measure P_0 on Ω (hence, it belongs to the same $\mathcal{B}$-equivalence class as P_0). Clearly, λ has density $d\lambda/d\lambda = 1$ with respect to itself, whereas it has density $d\lambda/d\mathsf{P}_0 = b - a$ w.r.t. P_0. Thus, a rescaling of the reference measure determines a rescaling of the *total*. For example, when λ is considered, the total is set to $\lambda(\Omega) = b - a$, whereas P_0 is associated with a total equal to $\mathsf{P}_0(\Omega) = 1$. On the other hand, once the scale of the reference measure is fixed, the corresponding densities satisfy the scale invariance property. For instance, having set the reference measure on $\Omega = [a,b]$ to λ, the Lebesgue density $d\lambda/d\lambda$ and the uniform density $d\mathsf{P}_0/d\lambda = \frac{1}{b-a}$ are equivalent. This is further exemplified in Section 4.3 where a detailed simulation study is provided. As such, it will always be necessary to specify the total mass of the reference measure as this matters for the analysis.

Since a typical choice for P is the Lebesgue measure, restricted here to a bounded support, it opens a question on how to change the reference from λ to a measure P with strictly positive λ-density $p = d\mathsf{P}/d\lambda$. This is done by using the well-known chain rule, i.e. for a generic measure μ we have that

$$\mu(\mathrm{B}) = \int_{\mathrm{B}} \frac{d\mu}{d\lambda}\, d\lambda = \int_{\mathrm{B}} \frac{d\mu}{d\lambda} \cdot \frac{d\lambda}{d\mathsf{P}}\, d\mathsf{P} = \int_{\mathrm{B}} \frac{d\mu}{d\lambda} \cdot \frac{1}{p}\, d\mathsf{P}.$$

Given a σ-finite measure P, the Bayes space $\mathcal{B}^2(\mathsf{P})$ is a space of $\mathcal{B}$-equivalence classes of σ-finite positive measures μ with square-integrable log-density w.r.t. the reference measure P:

$$\mathcal{B}^2(\mathsf{P}) = \left\{ \mu \in \mathcal{B}^2(\mathsf{P}) : \int \left| \ln \frac{d\mu}{d\mathsf{P}} \right|^2 d\,\mathsf{P} < +\infty \right\}, \tag{52}$$

where measures are identified with the corresponding Radon-Nikodym densities; or, equivalently, $\mathcal{B}^2(\mathsf{P})$ consists of $\mathcal{B}(\mathsf{P})$-equivalence classes of proportional density functions $f = \frac{d\mu}{d\mathsf{P}}$ on $\Omega = [a, b]$ whose logarithm is square-integrable w.r.t. P.

The reason for adopting a different reference measure P can be motivated by weighting itself, but it should be also remarked that it is necessary when dealing with PDFs on possibly unbounded supports [33].

6.3.1 Hilbert structure of weighted Bayes spaces

In this section, the definition of basic operations, perturbation and powering, and inner product under the general reference measure P will be considered. While both operations remain formally unchanged when changing the reference measure, the weighting affects the inner product. Here also the absolute scale of reference measure P matters, which corresponds to volume of the space Ω. It is possible to express densities from the Bayes space in the L^2 space (with respect to reference measure P) using clr transformation. This, however, still leaves open the problem of how to express the weighted densities in an *unweighted* L^2 space. A possible solution is presented in the thesis, Section 4.2.

Using a reference measure P, in [32] the operations of *perturbation* and *powering* are defined as

$$(\mu \oplus_{\mathsf{P}} \nu)(\mathrm{B}) =_{\mathcal{B}(\mathsf{P})} \int_{\mathrm{B}} \frac{d\mu}{d\mathsf{P}}(t) \cdot \frac{d\nu}{d\mathsf{P}}(t)\, d\mathsf{P}(t), \quad \mathrm{B} \in B \tag{53}$$

and

$$(\alpha \odot_{\mathsf{P}} \mu)(\mathrm{B}) =_{\mathcal{B}(\mathsf{P})} \int_{\mathrm{B}} \left(\frac{d\mu}{d\mathsf{P}}(t) \right)^{\alpha} d\mathsf{P}(t), \quad \mathrm{B} \in B, \tag{54}$$

where μ and ν are measures in $\mathcal{B}^2(\mathsf{P})$ and α is a real number. Moreover, all the measures μ, ν, λ and P are assumed to be well-defined. Consequently, these operations define a vector space structure on $\mathcal{B}^2(\mathsf{P})$ [32].

The operations (53) and (54) can be equivalently expressed using the densities with respect to P. Denoting them by $f_{\mathsf{P}} = \frac{d\mu}{d\mathsf{P}}$ and $g_{\mathsf{P}} = \frac{d\nu}{d\mathsf{P}}$ respectively, we have that

$$(f_{\mathsf{P}} \oplus_{\mathsf{P}} g_{\mathsf{P}})(t) =_{\mathcal{B}(\mathsf{P})} f_{\mathsf{P}}(t) \cdot g_{\mathsf{P}}(t) \quad \text{and} \quad (\alpha \odot_{\mathsf{P}} f_{\mathsf{P}})(t) =_{\mathcal{B}(\mathsf{P})} f_{\mathsf{P}}(t)^{\alpha}.$$

It is easy to verify that scale invariance of the reference density p holds for these operations. On the other hand, the scale of p is crucial for the definition of the inner product, defined originally in [33] and redefined here for the purpose of further developments as

$$\begin{aligned} \langle f_{\mathsf{P}}, g_{\mathsf{P}} \rangle_{\mathcal{B}(\mathsf{P})} &= \frac{1}{2\mathsf{P}(\Omega)} \int_{\Omega} \int_{\Omega} \ln \frac{f_{\mathsf{P}}(t)}{f_{\mathsf{P}}(u)} \ln \frac{g_{\mathsf{P}}(t)}{g_{\mathsf{P}}(u)} \, d\mathsf{P}(t) d\mathsf{P}(u) \\ &= \frac{1}{2\mathsf{P}(\Omega)} \int_{\Omega} \int_{\Omega} \ln \frac{f(t)}{f(u)} \ln \frac{g(t)}{g(u)} \cdot p(t) \cdot p(u) \, d\lambda(t) d\lambda(u), \end{aligned} \tag{55}$$

which endows the Bayes space $\mathcal{B}^2(\mathsf{P})$ with a separable Hilbert space structure. As a consequence, the distance between two densities $f_{\mathsf{P}}, g_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$ is obtained as

$$d_{\mathcal{B}(\mathsf{P})}(f_{\mathsf{P}}, g_{\mathsf{P}}) = \sqrt{\frac{1}{2\mathsf{P}(\Omega)} \int_{\Omega} \int_{\Omega} \left(\ln \frac{f_{\mathsf{P}}(t)}{f_{\mathsf{P}}(u)} - \ln \frac{g_{\mathsf{P}}(t)}{g_{\mathsf{P}}(u)} \right)^2 d\mathsf{P}(t) d\mathsf{P}(u).} \tag{56}$$

The reason for redefining the inner product (55) and distance (56) with respect to [33] reflects the approach presented in the multivariate case by Egozcue & Pawlowsky-Glahn [9], where the aim was to keep dominance under change of reference measure. Specifically, let p_0 be a uniform density of a measure P_0, not necessarily normalized to $\mathsf{P}_0(\Omega) = 1$, supported in an interval (or compact set) I in $\mathbb{R}$ (or $\mathbb{R}^m$), such that

$$\mathsf{P}_0(I) = \int_I p_0(t) \, dt < +\infty.$$

Let p, q be densities in $\mathcal{B}^2(\mathsf{P}_0)$ corresponding to measures P, Q such that P dominates Q, $\mathsf{P} \succ \mathsf{Q}$, that is

$$\mathsf{P}_0(t \in I : p(t) \geq q(t)) = \mathsf{P}_0(I).$$

Then, for $f_{\mathsf{P}_0}, g_{\mathsf{P}_0} \in \mathcal{B}^2(\mathsf{P}_0)$,

$$d_{\mathcal{B}(\mathsf{P})}(f_\mathsf{P}, g_\mathsf{P}) \geq d_{\mathcal{B}(\mathsf{Q})}(f_\mathsf{Q}, g_\mathsf{Q}), \tag{57}$$

where $f_\mathsf{P} = f_{\mathsf{P}_0} \cdot d\mathsf{P}_0/d\mathsf{P} =_{\mathcal{B}(\mathsf{P})} f_{\mathsf{P}_0} \ominus p_{\mathsf{P}_0}$ and $g_\mathsf{P} = g_{\mathsf{P}_0} \cdot d\mathsf{P}_0/d\mathsf{P} =_{\mathcal{B}(\mathsf{P})} g_{\mathsf{P}_0} \ominus p_{\mathsf{P}_0}$ [14]. The property (57) represents indeed the continuous counterpart to the subcompositional dominance in compositions [24]. That is, if the volume of the space $\mathsf{P}(I)$ is greater than or equal to $\mathsf{Q}(I)$ uniformly for any subinterval of I, then distances in $\mathcal{B}(\mathsf{P})$ dominate distances in $\mathcal{B}(\mathsf{Q})$. An example of this is comparing distances in a subinterval $I_1 \subseteq I$ with those in I – restrictions to subinterval corresponding to taking subcompositions [9].

Let's denote by $L_0^2(\mathsf{P})$ the closed subspace of $L^2(\mathsf{P})$ whose elements f_0 have zero integral $\int_\Omega f_0 \, d\mathsf{P} = 0$. Since the Bayes space $\mathcal{B}^2(\mathsf{P})$ is a Hilbert space, we can define an isometric isomorphism (i.e. a bijective map preserving distances) between $\mathcal{B}^2(\mathsf{P})$ and $L_0^2(\mathsf{P})$. Such a map is provided by the *centred logratio (clr)* transformation with respect to P, which is denoted by clr_P and is defined for $f_\mathsf{P} \in \mathcal{B}^2(\mathsf{P})$ by [33] as

$$f_\mathsf{P}^c(t) = \mathrm{clr}_\mathsf{P}(f_\mathsf{P})(t) = \ln f_\mathsf{P}(t) - \frac{1}{\mathsf{P}(\Omega)} \int_\Omega \ln f_\mathsf{P}(u) \, d\mathsf{P}(u), \quad t \in \Omega. \tag{58}$$

Its inverse mapping to $\mathcal{B}^2(\mathsf{P})$ is obtained by using the exponential transformation, $\exp[f_\mathsf{P}^c](t) = \exp[\mathrm{clr}_\mathsf{P}(f_\mathsf{P})](t)$, as shown in [33]. The clr representation allows to use the ordinary geometry of $L^2(\mathsf{P})$ to conduct operations of perturbation, powering, and inner product for the elements of $\mathcal{B}^2(\mathsf{P})$, while accounting for the specific features captured by the Bayes space. Indeed,

$$\mathrm{clr}_\mathsf{P}(f_\mathsf{P} \oplus_\mathsf{P} g_\mathsf{P}) = \mathrm{clr}_\mathsf{P}(f_\mathsf{P}) + \mathrm{clr}_\mathsf{P}(g_\mathsf{P}), \quad \mathrm{clr}_\mathsf{P}(\alpha \odot_\mathsf{P} f_\mathsf{P}) = \alpha \cdot \mathrm{clr}_\mathsf{P}(f_\mathsf{P})(t)$$

and

$$\langle f_\mathsf{P}, g_\mathsf{P} \rangle_{\mathcal{B}^2(\mathsf{P})} = \langle \mathrm{clr}_\mathsf{P}(f_\mathsf{P}), \mathrm{clr}_\mathsf{P}(f_\mathsf{P}) \rangle_{L^2(\mathsf{P})} . \tag{59}$$

Unlike the case of [33, Sect. 4], in this work the reference measure P in $L_0^2(\mathsf{P})$ is not necessarily a probability measure, as its normalization may lead to incoherent results when restricting the analysis to a subdomain of the original domain Ω (as was shown in the discrete case [9]).

6.3.2 Unweighting Bayes spaces

Most methods developed for FDA rely on the assumption that functional data are embedded in the *unweighted* L^2 space. However, the clr transformation (58) maps measures/densities in (a subspace of) a weighted space L^2 space, i.e. $L_0^2(\mathsf{P})$. Similarly, methods developed so far in Bayes spaces ground on the assumption that a uniform reference measure is considered, as for instance in Sections 2.2 and 3. A transformation mapping P-densities from $\mathcal{B}^2(\mathsf{P})$ to an unweighted counterpart of $L_0^2(\mathsf{P})$ would have the advantage of allowing the use of most FDA methods while accounting for the weighted Bayes structure of the data. Similarly, a transformation mapping P-densities from $\mathcal{B}^2(\mathsf{P})$ to an *unweighted* space $\mathcal{B}^2(\lambda)$ would allow for the use of unweighted methods to perform actual computations. In this subsection, we derive an unweighting scheme allowing one to represent the weighted Bayes space geometry in an unweighted Bayes space, as well as in an unweighted L^2 space.

We thus aim to define three mappings. Firstly, we define ω from $\mathcal{B}^2(\lambda)$ to $\mathcal{B}^2(\mathsf{P})$ as a *weighting* map associating an unweighted λ-density to a weighted P-density. Inversely, ω^{-1} is interpreted as an *unweighting* map. Similarly, we define ω_2 and its inverse ω_2^{-1} which play the same role between the unweighted and weighted L^2 spaces, i.e. $L^2(\lambda)$ and $L^2(\mathsf{P})$ respectively. Finally, we define clr_u (*unweighting clr*) such that, for $f_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$,

$$\mathrm{clr}_u(f_{\mathsf{P}} \oplus_{\mathsf{P}} g_{\mathsf{P}}) = \mathrm{clr}_u(f_{\mathsf{P}}) + \mathrm{clr}_u(g_{\mathsf{P}}), \quad \mathrm{clr}_u(\alpha \odot f_{\mathsf{P}}) = \alpha \cdot \mathrm{clr}_u(f_{\mathsf{P}})(t)$$

and

$$\langle f_{\mathsf{P}}, g_{\mathsf{P}} \rangle_{\mathcal{B}^2(\mathsf{P})} = \langle \mathrm{clr}_{\mathsf{u}}(f_{\mathsf{P}}), \mathrm{clr}_{\mathsf{u}}(f_{\mathsf{P}}) \rangle_{L^2(\lambda)} . \tag{60}$$

To support this construction and study the properties of these maps, we shall use

$$\begin{tikzcd}
L^2_0(\lambda) \arrow[r, shift left, leftarrow, "\mathrm{clr}_\lambda"] \arrow[r, shift right, "\exp"'] & \mathcal{B}^2(\lambda) \arrow[r, shift left, "\ominus p"] \arrow[r, shift right, leftarrow, "\oplus p"'] & \mathcal{B}^2(\mathsf{P}) \arrow[r, shift left, "\mathrm{clr}_{\mathsf{P}}"] \arrow[r, shift right, leftarrow, "\exp"'] \arrow[d, shift left, "\omega^{-1}"] \arrow[d, shift right, leftarrow, "\omega"'] \arrow[dr, "\mathrm{clr}_u"] & L^2_0(\mathsf{P}) \arrow[d, shift left, "\omega_2^{-1}"] \arrow[d, shift right, leftarrow, "\omega_2"'] \\
 & & \mathcal{B}^2(\lambda) \arrow[r, shift left, "\mathrm{clr}_{\sqrt{\mathsf{P}}}"] \arrow[r, shift right, leftarrow, "\exp"'] & L^2_{0,\sqrt{\mathsf{P}}}(\lambda)
\end{tikzcd}$$

Figure 1: Relationships among weighted and unweighted Bayes spaces, $\mathcal{B}^2(\mathsf{P})$ and $\mathcal{B}^2(\lambda)$, and weighted and unweighted $L^2(\mathsf{P})$ and $L^2(\lambda)$ spaces.

an auxiliary measure $\sqrt{\mathsf{P}}$ defined as

$$\sqrt{\mathsf{P}}(\mathrm{A}) = \int_{\mathrm{A}} \sqrt{p}\, d\lambda, \quad \mathrm{A} \in \mathcal{A}.$$

This measure plays the role of *unweighting* measure, in the sense that it allows to consistently map the *weighted* Bayes space $\mathcal{B}^2(\mathsf{P})$ into a subset of the *unweighted* L^2 space. We refer the reader to the scheme in Figure 1 as a concise representation of these relationships.

We define the $\mathcal{B}^2$-*weighting* map ω as

$$\begin{aligned} \omega : \mathcal{B}^2(\lambda) &\to \mathcal{B}^2(\mathsf{P}) \\ \varphi &\mapsto \omega(\varphi) = \varphi^{1/\sqrt{p}}, \end{aligned} \tag{61}$$

where $p = \frac{d\mathsf{P}}{d\lambda}$ (recall: p is assumed to be strictly positive in Ω). In (61), the map ω is formulated for measures, but it can be equivalently expressed using densities with respect to respective reference measures. This map defines a bijection between $\mathcal{B}^2(\lambda)$ and $\mathcal{B}^2(\mathsf{P})$, as proved in the following proposition.

Proposition 6.3.1 *The map ω defined in (61) is one-to-one and onto.*

The inverse ω^{-1} is defined as $\omega^{-1}(\psi) = \psi^{\sqrt{p}}$ and it is interpreted as a $\mathcal{B}^2$-*unweighting* map. It is represented in the bottom left part of the scheme in Figure 1. Obviously, both ω and ω^{-1} depend on the scale of P.

We define the L^2-*weighting* map ω_2 as

$$\begin{aligned} \omega_2 : L^2(\lambda) &\to L^2(\mathsf{P}) \\ \eta &\mapsto \omega(\eta) = \eta/\sqrt{p}. \end{aligned}$$

Using the same rationale as for Proposition 6.3.1, it can be proved that ω_2 defines a bijection between $L^2(\lambda)$ and $L^2(\mathsf{P})$. Its inverse ω_2^{-1} is defined as $\omega_2^{-1}(\xi) = \xi\sqrt{p}$ and it is interpreted as a L^2-*unweighting* map. It is represented in the bottom right part of the scheme in Figure 1. Note that ω is non-linear with respect to the Bayes space geometry, as well as ω_2 is non-linear in L^2.

Using (55), the map $\mathrm{clr}_u : \mathcal{B}^2(\mathsf{P}) \to L^2(\lambda)$ can be then defined as

$$\mathrm{clr}_u(f_{\mathsf{P}}) = \omega_2^{-1}[\mathrm{clr}_{\mathsf{P}}(f_{\mathsf{P}})]. \tag{62}$$

It can be proven that (62) fulfills all the properties detailed in (60). Note that the scale of clr_u depends on the scale of $\sqrt{p}$, hence on the scale of $\sqrt{\mathsf{P}}$, because of the non-linearity of ω_2 (see [6] for the case of finite-dimensional compositions). As such, similarly to the multivariate case [9], the scale of the reference measure is relevant in the geometry of both weighted and unweighted spaces.

It is worth noticing that clr_u is closely related to a different centered logratio transformation. This is defined on the unweighted space $\mathcal{B}^2(\lambda)$ and induced by the unweighting measure $\sqrt{\mathsf{P}}$. Indeed, let $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ be the subspace of $L^2(\lambda)$ such that $\int_\Omega f \, d\sqrt{\mathsf{P}} = 0$ for $f \in L^2(\lambda)$. Let's define on $\mathcal{B}^2(\lambda)$ the map $\mathrm{clr}_{\sqrt{\mathsf{P}}}$ as

$$\mathrm{clr}_{\sqrt{\mathsf{P}}}(\varphi)(t) = \ln \varphi(t) - \frac{1}{\sqrt{\mathsf{P}}(\Omega)} \int_\Omega \ln[\varphi(u)] \, d\sqrt{\mathsf{P}}(u), \quad t \in \Omega, \quad \varphi \in \mathcal{B}^2(\lambda). \tag{63}$$

In light of Proposition 6.3.1, it is easy to see that the map (63) is well defined. For any $\varphi \in \mathcal{B}^2(\lambda)$, we can set $f_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$ to $f_{\mathsf{P}} = \omega(\varphi) = \varphi^{1/\sqrt{p}}$. Then, it holds that

$$\int_\Omega \ln[\varphi(u)] \, d\sqrt{\mathsf{P}}(u) = \int_\Omega \ln[f_{\mathsf{P}}(u)] p(u) \, d\lambda(u) < +\infty.$$

Moreover, for any φ in $\mathcal{B}^2(\lambda)$, we have that $\mathrm{clr}_{\sqrt{\mathsf{P}}}(\varphi) \in L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$. The following proposition establishes the close relationship between clr_u and $\mathrm{clr}_{\sqrt{\mathsf{P}}}$, thus completing the scheme in Figure 1.

Proposition 6.3.2 *The following statements hold true.*

(i) The image of the space $\mathcal{B}^2(\mathsf{P})$ *under the map* clr_u *defined in* (62) *is* $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$.

(ii) The map clr_u *coincides with the composed function* $\mathrm{clr}_{\sqrt{\mathsf{P}}} \circ \omega^{-1}$, *i.e.*

$$\mathrm{clr}_u(f_{\mathsf{P}}) = \mathrm{clr}_{\sqrt{\mathsf{P}}}(\omega^{-1}(f_{\mathsf{P}})) \quad \textit{and} \quad f_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P}).$$

(iii) The inverse of the map $\mathrm{clr}_{\sqrt{\mathsf{P}}}$ *is* $\mathrm{clr}^{-1}_{\sqrt{\mathsf{P}}} : L^2_{0,\sqrt{\mathsf{P}}}(\lambda) \to \mathcal{B}^2(\lambda)$ *and is given by*

$$\mathrm{clr}^{-1}_{\sqrt{\mathsf{P}}}(\psi) =_{\mathcal{B}^2(\lambda)} \exp(\psi),$$

for any ψ *in* $L^2_{0,\sqrt{\mathsf{P}}}$.

(iv) The inverse of the map clr_u *is* $\mathrm{clr}^{-1}_u : L^2_{0,\sqrt{\mathsf{P}}}(\lambda) \to \mathcal{B}^2(\mathsf{P})$ *and is given by*

$$\mathrm{clr}^{-1}_u(\psi) =_{\mathcal{B}^2(\mathsf{P})} \exp[\omega_2(\psi)] =_{\mathcal{B}^2(\mathsf{P})} \omega[\exp(\psi)],$$

for any ψ *in* $L^2_{0,\sqrt{\mathsf{P}}}$.

Note that taking the $\mathcal{B}^2$-*unweighting* transformation ω^{-1} is indeed different from simply changing the reference measure from P to λ. The former transformation is indeed used to *represent* the weighted Bayes space through an unweighted one, while preserving its weighted Hilbert geometry. In fact, as further highlighted in Section 5.2, this auxiliary space may serve to enhance the interpretation of the weighted structure. For instance, visual interpretation of a weighted density f_{P} in $\mathcal{B}^2(\mathsf{P})$ is hindered by the need to take into account the weighting scheme considered for the support. On the contrary, visualisation of the corresponding unweighted density $\omega^{-1}(f_{\mathsf{P}})$ allows for the usual interpretation, yet representing the same object – just by incorporating the weighting scheme.

It is also clear that, as long as the Lebesgue reference measure is concerned ($\mathsf{P}(\Omega) = \lambda([a, b])$), the transformations clr_u and $\mathrm{clr}_{\mathsf{P}}$ coincide, and they reduce to the clr transformation clr_λ (8). Note, however, that this would not be true for reference measures proportional to the Lebesgue one, because the scale of the reference does have an impact on the Hilbert geometry.

The above considerations have a direct impact on applications. For a sample of densities $f_1, \dots, f_N$ to be analyzed with respect to a reference measure P, the following strategy can be adopted:

1. Set the reference measure P.
2. If the PDFs were given w.r.t. the Lebesgue measure, change the reference measure from λ to P. That is, set $f_{\mathsf{P},i} = f_i \ominus p$, for $i = 1, \dots, N$, with $f_{\mathsf{P},i} \in \mathcal{B}^2(\mathsf{P})$.
3. Map $f_{\mathsf{P},i}$, for $i = 1, \dots, N$, onto $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ by using the clr_u transformation. Set $y_i = \mathrm{clr}_u(f_{\mathsf{P},i})$, for $i = 1, \dots, N$.
4. Perform the statistical analysis on y_i, $i = 1, \dots, N$, using *unweighted* L^2_0 ($L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$) methods.
5. If the results needs to be given in terms of densities, use the inverse transformation $\exp[\mathrm{clr}_u(f_{\mathsf{P}})]$ to express the results in the unweighted space $\mathcal{B}^2(\lambda)$, where they can be easily interpreted.

This strategy is further illustrated in the Section 5, which presents a dimensionality reduction method in weighted Bayes spaces.

6.4 Statistical methods in weighted Bayes spaces: weighted SFPCA

Simplicial functional principal component analysis (SFPCA, [15]) was recently introduced to adapt the well-known functional principal component analysis [27] to density functions. It is grounded on the theory of Bayes spaces and assumes that the Lebesgue measure is set as a reference measure. SFPCA aims to explore the main modes of *relative* variability in a sample of density data and can be used to suggest a possible dimensionality reduction of a dataset of PDFs. In this section, we extend the SFPCA to its weighted version, named hereafter wSFPCA. Besides its relevance in applications, this extension serves as an illustrative example of the strategy detailed in Section 4.2.

Let's denote by $f_1, \dots, f_N$ an i.i.d. sample in $\mathcal{B}^2(\lambda)$. After selecting the reference measure P with λ-density p, a sample $f_{\mathsf{P},i} = f_i \ominus p$, for $i = 1, \dots, N$, in $\mathcal{B}^2(\mathsf{P})$ is obtained. We assume without loss of generality this sample is mean-centered. If this is not the case, it is enough to consider $\tilde{f}_{\mathsf{P},i} = f_{\mathsf{P},i} \ominus \bar{f}_{\mathsf{P}}$, where $\bar{f}_{\mathsf{P}}$ stands for the (weighted) sample mean of the observed (weighted) densities

$$\bar{f}_{\mathsf{P}} = \frac{1}{N} \odot_{\mathsf{P}} \bigoplus\nolimits_{\mathsf{P}_{i=1}}^{N} f_{\mathsf{P},i}.$$

Note that the centering operation shifts the center of the sample to the neutral element of the (weighted) perturbation operation, that is, the uniform density on $\mathcal{B}^2(\mathsf{P})$.

The aim of wSFPCA is to identify a collection of orthogonal and normalized P-density functions $\{\xi_{\mathsf{P},j}\}_{j\geq 1}$ in $\mathcal{B}^2(\mathsf{P})$ corresponding to the directions in $\mathcal{B}^2(\mathsf{P})$ along which the dataset displays its main modes of variability. These directions are called weighted simplicial functional principal components (wSFPCs), and they are obtained by maximizing the following objective function

$$\sum_{i=1}^{N} \langle f_{\mathsf{P},i}, \xi_{\mathsf{P}} \rangle^2_{\mathcal{B}(\mathsf{P})} \text{ subject to } \|\xi_{\mathsf{P}}\|_{\mathcal{B}(\mathsf{P})} = 1; \text{ with } \qquad \langle \xi_{\mathsf{P}}, \xi_{\mathsf{P},k} \rangle_{\mathcal{B}(\mathsf{P})} = 0, \, k < j, \tag{64}$$

over ξ_{P} in $\mathcal{B}^2(\mathsf{P})$, where $\langle f_{\mathsf{P},i}, \xi_{\mathsf{P}} \rangle_{\mathcal{B}(\mathsf{P})}$ is the projection of $f_{\mathsf{P},i}$ along the direction in $\mathcal{B}^2(\mathsf{P})$ identified by ξ_{P}, i.e., coordinate of f_{P} (Fourier coefficient). The orthogonality condition has only to be fulfilled for $j \geq 2$, and guarantees that the jth wSFPC $\xi_{\mathsf{P},j}$ is orthogonal to the first $j-1$ wSFPCs.

Since $\mathcal{B}^2(\mathsf{P})$ is a Hilbert space, the solution of the maximization problem (64) exists and is unique for all $j \in \{1, 2, \dots, N-1\}$. It coincides with the set of eigenfunctions associated with the ordered eigenvalues of the sample covariance operator $V : \mathcal{B}^2(\mathsf{P}) \to \mathcal{B}^2(\mathsf{P})$, defined for $\xi_{\mathsf{P}} \in \mathcal{B}^2(\mathsf{P})$ as

$$V\xi_{\mathsf{P}} = \frac{1}{N} \odot_{\mathsf{P}} \bigoplus\nolimits_{\mathsf{P}_{i=1}}^{N} \langle f_{\mathsf{P},i}, \xi_{\mathsf{P}} \rangle_{\mathcal{B}(\mathsf{P})} \odot_{\mathsf{P}} f_{\mathsf{P},i}. \tag{65}$$

The jth wSFPC $\xi_{\mathsf{P},j}$ is thus obtained by solving the eigenequation $V\xi_{\mathsf{P},j} = \rho_j \odot_{\mathsf{P}} \xi_{\mathsf{P},j}$. The $N-1$ eigenvalues $\rho_1 \geq \dots \geq \rho_{N-1}$ represent the variability of the dataset along the directions of the associated eigenfunctions $\xi_{\mathsf{P},1}, \dots, \xi_{\mathsf{P},N-1}$.

From the practical viewpoint, it is desirable to restate the problem of finding the eigenpairs $(\xi_{\mathsf{P},j}, \rho_j), j = 1, \ldots, N-1$, in $\mathcal{B}^2(\mathsf{P})$ in terms of the unweighted L^2 spaces, i.e. $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$, where well-established computational methods are available. To this end, consider the clr_u transformation of the data, i.e. $\mathrm{clr}_u(f_{\mathsf{P},1}), \ldots, \mathrm{clr}_u(f_{\mathsf{P},N})$. Following the same arguments of [15], one can easily prove that performing a functional principal component analysis of the transformed dataset in $L^2_{0,\sqrt{\mathsf{P}}}(\lambda)$ yields the eigenpairs $\left(\mathrm{clr}_u(\xi_{\mathsf{P},j}), \rho_j\right), j = 1, \ldots, N-1$. The resulting eigenfunctions $\mathrm{clr}_u(\xi_{\mathsf{P},j})$ can be eventually transformed back into $\mathcal{B}^2(\mathsf{P})$, or into the unweighted $\mathcal{B}^2(\lambda)$, by using the corresponding inverse clr transformation (i.e. clr_u^{-1} or $\mathrm{clr}_{\sqrt{\mathsf{P}}}^{-1}$ respectively) to proceed with interpretation in the original space.

The results of wSFPCA can be interpreted, e.g. by analyzing the principal component scores, which are useful to inspect the relationships among observations. Note that the score f_{ij} is a projection of the (centered) observation $f_{\mathsf{P},i}$ along the direction $\xi_{\mathsf{P},j}$, i.e. $f_{ij} = \langle f_{\mathsf{P},i}, \xi_{\mathsf{P},j} \rangle_{\mathcal{B}(\mathsf{P})} = \langle \mathrm{clr}_u(f_{\mathsf{P},i}), \mathrm{clr}_u(\xi_{\mathsf{P},j}) \rangle_{L^2(\lambda)}$, and thus the scores coincide in $\mathcal{B}^2(\mathsf{P})$ and $L^2(\lambda)$. It is useful to visualize the mean density perturbed by the jth wSFPC $\xi_{\mathsf{P},j}$ powered by a suitable coefficient. This represents the variability around the mean function along the direction of a given wSFPC, and can support the analyst in the definition of a weighting strategy for the dataset at hand. Indeed, in the context of general reference measures, the wSFPCs can be plotted and interpreted to see the effect of weighting the domain of the distributional variable according to alternative reference measures. Finally, for the purpose of dimensionality reduction, the number of wSFPCs to be retained can be set by the commonly used scree plot. Particularly, searching for an elbow shape or setting a threshold on the portion of variance explained by wSFPCs as usually.

7 Original results and summary

The focus of this work was to develop statistical methods for the analysis of functional data carrying relative information – probability density functions, defined on a bounded domain. These methods are grounded on the theory of Bayes Hilbert spaces, capturing all key inherent features of densities (i.e., scale invariance, relative scale), and they extend the well-known results of FDA to density functions.

In Section 2, we considered the problem of statistical preprocessing of densities using spline functions, performed in the clr space. Firstly, we recalled optimal smoothing splines for clr transformed density functions as proposed in [17]. Here, we proved a new key result to characterize B-spline representation of clr transformed densities using standard B-spline basis system in terms of a linear constraint on the B-spline basis coefficients. Nevertheless, it was recognized that using the standard B-spline basis system for approximation of density functions in clr space has some limitations since the basis elements do not belong to the L^2 space. Therefore, this approach was updated by proposing a new class of compositional splines which enable to construct a B-spline basis directly in the clr space of density functions (ZB-spline basis system) and, consequently, also in the original space of densities (CB-spline basis system). Accordingly, compositional splines can be implemented instead of the standard ones into FDA methods for statistical processing of density functions. Also further tuning of the compositional splines is possible, here represented by the smoothing compositional splines or by orthonormalization of the ZB-basis systems. As for future research, it could be attractive to generalize the methodology of compositional splines even for multidimensional density functions.

In Section 3, a novel approach to perform functional regression when the response is a density function using the Bayes space methodology was developed. For the actual estimation of the regression coefficients, an approach based on B-spline expansion of clr transformed density functions was proposed. This expansion enables to control the smoothness of the estimated regression coefficients (density functions) through the smoothness of the B-spline representation of the response. On the other hand, it turned out that the linear constraint on B-spline basis coefficients (using the former approach for B-spline expansion of

PFDs) induces the singularity problem into the regression model. Nevertheless, this can be overcome by using the compositional splines which lead to expression of PDFs through a set of unconstrained coefficients. Such representation can be then further used for the purpose of inference on the coefficients using proper functional tests.

The role of reference measure in Bayes spaces was discussed in Section 4, specially, a novel weighting approach to probability density functions was proposed. An advanced weighting scheme was developed which enables to link weighted Bayes spaces to unweighted $\mathcal{B}^2$ and L^2 spaces. The advantage of representing weighted densities in an unweighted space is demonstrated by the possibility of (i) making comparisons of densities arising from different weighting criteria, and (ii) visually interpret the results through ordinary 'unweighted eyes'. In fact, the proposed framework allows to perform statistical processing in weighted Bayes spaces by using simply popular (unweighted) methods, which were developed for FDA. In the final Section 5, this strategy has been demonstrated by extending a dimensionality reduction method (SFPCA) to the weighted case. Nevertheless, other methods could be considered as well, such as clustering, regression, spatial prediction techniques, etc. We finally stress that considering different weighting schemes can be particularly relevant in statistical applications, i.e., (i) to account for different degrees of uncertainty across the domain of the data, (ii) to incorporate prior knowledge about the phenomenon or (iii) to perform domain selection.

I truly hope that the presented thesis helps to expand the Bayes space methodology for statistical processing of density functions and that it will be a motivation to propose other statistical methods for analyzing PDFs such as outlier detection and related anomaly detection, classification or functional regression with densities playing the role of the response and/or covariates.

List of publications

Research papers

- M. A. Álvarez-Vázquez, M. Hošek, J. Elznicová, J. Pacina, K. Hron, K. Fačevicová, **R. Talská**, and O. Bábek, Separation of geochemical signals in fluvial sediments: new approaches to grain size control and element contamination (*submitted*).

- M. Hošek, J. Pacina, J. Štojdl, O. Bábek, J. Sedláček, K. Hron, **R. Talská**, S. Křížinecká, J. Fikarová, and T. Matys Grygar, Change in geochemistry of fluvial sediments after dam construction (the Chrudimka River, the Czech Republic). *Applied Geochemistry*, 98:94-108, 2018.

- J. Machalová, **R. Talská**, K. Hron, and A. Gába, Compositional splines for representation of density functions (*under review*).

- **R. Talská**, J. Machalová, P. Smýkal, and K. Hron, A comparison of seed germination coefficients using functional regression (*Applications in Plant Sciences, accepted for publication*).

- **R. Talská,**, A. Menafoglio, K. Hron, J. J. Egozcue, and J. Palarea-Albaladejo, Weighting the domain of probability densities in functional data analysis (*Stat, accepted for publication*).

- **R. Talská**, A. Menafoglio, J. Machalová, K. Hron, and E. Fišerová, Compositional regression with functional response. *Computational Statistics and Data Analysis*, 123:66-85, 2018.

List of conferences

- AMISTAT, 10.-13.11.2016, Prague (CZ): Exponential families from the perspective of Bayes spaces: A simulation study (poster)

- COMPSTAT, 23.-26.8.2016, Oviedo (ES): Functional regression analysis with compositional response (presentation)

- ROBUST, 11.-16.9.2016, Jeseníky (CZ): Kompoziční regrese s funkcionální závisle proměnnou (poster+presentation, in Czech)

- ERCIM, 9.-11.12.2016, Seville (ES): Functional regression analysis with compositional response (presentation)

- ODAM, 31.5.-2.6.2017, Olomouc (CZ): Linear regression models with distributional response (presentation)

- CoDaWork 2017, 5.6.-9.6.2017, Abbadia San Salvatore (IT): Functional linear regression models with distributional response (presentation)

- ERCIM, 16.-18.12.2017, London (GB): Effects of changing the reference measure in statistical processing of density functions (presentation)

- DSSV 2018, 9.-11.7.2018, Vienna (AT): Changing the reference measure and its effects in statistical processing of density data (presentation)

- CRoNoS, 14.-16. 4. 2019, Limassol (CY): Weighting in Bayes spaces and its effects in statistical processing of density functions (presentation)

- ODAM, 29.-31.5.2019, Olomouc (CZ): Weighted Bayes spaces (presentation)

- CoDaWork 2019, 3.–7.6.2019, Terrassa (ES): Changing reference measure in Bayes spaces and its effect in statistical processing of density functions (presentation)

References

[1] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177, 1982.

[2] R. S. Bivand, J. Wilk, and T. Kossowski. Spatial association of population pyramids across Europe: The application of symbolic data, cluster analysis and join-count tests. *Spatial Statistics*, 21:339–361, 2017.

[3] C. De Boor. *A practical guide to splines*, Springer-Verlag, New York, 1978.

[4] P. Dierckx. *Curve and surface fitting with splines.* Oxford University Press, Oxford, 1995.

[5] J. J. Egozcue, V. Pawlowsky-Glahn, R. Tolosana-Delgado, M. Ortego, and K. van den Boogaart. Bayes spaces: use of improper distributions and exponential families. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 107(2):475–486, 2013.

[6] J. J. Egozcue, C. Barcelo-Vidal, J. Martín-Fernández, E. Jarauta-Bragulat, J. Díaz-Barrero, G. Mateu-Figueras, V. Pawlowsky-Glahn, and A. Buccianti. Elements of simplicial linear algebra and geometry. *Compositional data analysis: Theory and applications*, 139–157, 2011.

[7] J. J. Egozcue, P. Daunis-i Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser. Simplicial regression. The normal model. *Journal of Applied Probability and Statistics (JAPS)*, 6:87–108, 2012.

[8] J. J. Egozcue, J. L. Díaz-Barrero, and V. Pawlowsky-Glahn. Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*, 22(4):1175–1182, 2006.

[9] J. J. Egozcue and V. Pawlowsky-Glahn. Changing the reference measure in the simplex and its weighting effects. *Austrian Journal of Statistics*, 45(4):25–44, 2016.

[10] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[11] J. J. Faraway. Regression analysis for a functional response. *Technometrics*, 39(3):254–261, 1997.

[12] V. Ferro and S. Mirabile. Comparing particle size distribution analysis by sedimentation and laser diffraction method. *Journal of Agricultural Engineering*, 2:35–43, 2009.

[13] E. Fišerová, L. Kubáček, and P. Kunderová. *Linear Statistical Models: Regularity and Singularities*. Academia, 2007.

[14] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[15] K. Hron, A. Menafoglio, M. Templ, K. Hruzová, and P. Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, 94:330–350, 2016.

[16] P. Kokoszka, H. Miao, A. Petersen, and H. L. Shang. Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting*, 35(4):1304–1317, 2019.

[17] J. Machalová, K. Hron, and G. S. Monti. Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*, 43(8):1419–1435, 2016.

[18] J. Machalová, R. Talská, K. Hron, and A. Gába. Compositional splines for representation of density functions. *Under review*.

[19] G. Mateu-Figueras and V. Pawlowsky-Glahn. A critical approach to probability laws in geochemistry. *Progress in Geomathematics*, pages 39–52. Springer, 2008.

[20] A. Menafoglio, G. Gaetani, and P. Secchi. Random domain decompositions for object-oriented Kriging over complex domains. *Stochastic Environmental Research and Risk Assessment*, 32(12):3421–3437, 2018.

[21] A. Menafoglio, A. Guadagnini, and P. Secchi. A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment*, 28(7):1835–1851, 2014.

[22] A. Menafoglio, A. Guadagnini, and P. Secchi. Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a Bayes space approach. *Water Resources Research*, 52(8):5708–5726, 2016.

[23] A. Menafoglio, P. Secchi, and A. Guadagnini. A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. *Mathematical Geosciences*, 48(4):463–485, 2016.

[24] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. *Modeling and analysis of compositional data.* John Wiley & Sons, 2015.

[25] A. Petersen and H.-G. Müller. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1):183–218, 2016.

[26] A. Petersen and H.-G. Müller. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2):339–351, 2019.

[27] J. Ramsay and B. Silverman. Functional data analysis, second edition. *Springer Series in Statistics*, 2005.

[28] W.-K. Seo and B. Beare. Cointegrated linear processes in Bayes Hilbert space. *Statistics & Probability Letters*, 147:90–95, 2018.

[29] Q. Shen and H. Xu. Diagnostics for linear models with functional responses. *Technometrics*, 49(1):26–33, 2007.

[30] R. Talská, A. Menafoglio, K. Hron, J. J. Egozcue, and J. Palarea-Albaladejo. Weighting the domain of probability densities in functional data analysis. *Stat, accepted for publication.*

[31] R. Talská, A. Menafoglio, J. Machalová, K. Hron, and E. Fišerová. Compositional regression with functional response. *Computational Statistics & Data Analysis*, 123:66–85, 2018.

[32] K. G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes linear spaces. *SORT*, 34(4):201–222, 2010.

[33] K. G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194, 2014.