Akademie věd
České republiky

Teze disertace
k získání vědeckého titulu "doktor věd"
ve skupině věd fyzikálně-matematických

Density data analysis in Bayes spaces and its applications

Komise pro obhajoby doktorských disertací v oboru Informatika a kybernetika

Jméno uchazeče: prof. RNDr. Karel Hron, Ph.D.

Pracoviště uchazeče: Katedra matematické analýzy a aplikací matematiky,
Přírodovědecká fakulta, Univerzita Palackého v Olomouci

Místo a datum: Olomouc, 1.3.2025

Contents

Preface

Recognize your sample space for what it is. Pay attention to its properties and follow through any logical necessities arising from these properties. The solution here to the apparent awkwardness of the sample space is not so difficult. The difficulty is facing up to reality and not imagining that there is some esoteric panacea.

J. Aitchison: The one-hour course in compositional data analysis or compositional data analysis is simple. IAMG'97, Barcelona, 1997.

In 1982, John Aitchison published a seminal paper on the log-ratio approach to compositional data analysis (Aitchison, 1982), in which he recognised that these are essentially *scale invariant* observations. In other words, although compositional data have traditionally been defined as positive multivariate observations with the unit sum constraint, the relevant information is actually contained in the (log) ratios between their components. This mental step gave rise to the so-called log-ratio methodology of compositional data, which has now become a methodological mainstream in applications, but still one step further was needed to recognise that probability density functions are also essentially of the same nature (Egozcue et al., 2006). And much more, that there can be derived a general framework of Bayes spaces with the Hilbert space structure (van den Boogaart et al., 2014), which covers both compositional data and probability density functions (PDFs) as discrete and continuous *distributional* (relative) *data*. Bayes spaces opened up the possibility of considering PDFs as data objects. However, their statistical processing by popular methods of functional data analysis (Ramsay and Silverman, 2005) must be done carefully, because their sample space is formed by equivalence classes of proportional positive functions, far enough from the usual L^2 space. This was first demonstrated in Delicado (2011) and further developments over the years led to the current state of a rapidly growing community, which can be documented by recent publications in renowned journals (Lei et al., 2023; Eckardt et al., 2024; Ma et al., 2024; Murph et al., 2024; Qiu et al., 2024; Kutta et al., 2025). The potential of Bayes spaces was recognized also outside the core community: in the paper Petersen et al. (2022) whose aim was to compare different methodologies for modeling PDFs as data objects, the conclusion about looking beyond univariate densities was that "The Bayes

space representation, (...), provides a sound theoretical base for multivariate densities, although its practical implementation and utility has only been given limited, if any, consideration." Indeed, the generalisation of Bayes spaces to the multivariate setting was addressed in Genest et al. (2023) and its first concise application was presented in Matys Grygar et al. (2024).

I'm grateful to have the opportunity, with the great support of my colleagues, to contribute to many important developments in the field. The aim of this thesis is to summarise some of the most important of these. It consists of a general (and rather simplified) introduction to density data analysis and a specification of the Candidate's contribution to the research area, followed by a collection of scientific papers illustrating the corresponding publication activities. The aim of the thesis is thus to present developments in density data analysis using Bayes spaces, specifically from two aspects:

1. to build a concise methodology for density data processing by adapting popular methods of functional data analysis;

2. to contribute to the theoretical development of Bayes spaces themselves.

These two aspects are reflected in the papers that constitute the thesis.

Most of the work reported in the thesis is the result of extensive collaboration with my colleagues and PhD students. I would like to thank them all for the cooperation over the past years and look forward to further joint scientific activities.

Olomouc, February 2025

Karel Hron

1 Compositional and density data as relative data

Compositional data are traditionally presented in the literature as positive *constrained* multivariate data (Chayes, 1960; Scealy and Welsh, 2011; Tsagris and Stewart, 2018), which means that the sample space of D-part compositions $\mathbf{x} = (x_1, \ldots, x_D)$ is the simplex,

$$S^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbf{R}^D \mid x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa \right\}. \tag{1}$$

But already John Aitchison recognised in Aitchison (1986), Property 2.3 that the sample space of compositional data is in fact more general, consisting of equivalence classes of proportional positive vectors, i.e. vectors multiplied by a positive constant – and this idea was further developed in the following decades, leading to the current definition of compositional data as *scale invariant* objects (Pawlowsky-Glahn et al., 2015), where the actual representation of compositions is irrelevant for their analysis. Consequently, the simplex sample space (1) is the sample space of *unit sum representations* of multivariate data whose relevant information is contained in (log-)ratios between components (compositional parts). This *relative* information is fully captured by representing compositions in terms of *logcontrasts*, i.e. loglinear combinations

$$a_1 \ln x_1 + \ldots + a_D \ln x_D$$

with

$$a_1 + \ldots + a_D = 0.$$

With logcontrasts the *logratio methodology of compositional data* was born. It is essential to analyse compositions or related data in an appropriate sample space, for which the respective methods are designed, otherwise their statistical processing may lead to useless results (Pearson, 1897; Chayes, 1960; Aitchison, 1986; Filzmoser et al., 2018). While approaches that treat compositions as *constrained* data attempt, usually with considerable effort, to adapt popular statistical methods to the simplex, the strategy developed in the logratio methodology is different: to represent compositional data in an appropriate set of logcontrasts, and then to proceed with popular multivariate statistical methods (taking into account the interpretation of these logcontrasts) in the real space for which these methods were developed (Eaton,

1983). There is still an ongoing discussion about which choice of logcontrasts is the best (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018; Greenacre, 2018); in any case, the logratio methodology is now the mainstream approach for the statistical processing of compositions. However, one point was clear from the beginning: if the logratio methodology was to be developed beyond the presentation of practical and interpretable sets of logcontrasts for the analysis of compositions, the geometric structure of the sample space was needed. For compositional data, the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001; Billheimer et al., 2001) with the Euclidean vector space structure was developed, which fully reflects them as scale invariant observations. And allowed to generalise the pioneering and truly fundamental ideas of John Aitchison to other data objects carrying relative information, in particular to *probability density functions*.

Probability density functions (PDFs), on which we focus in the thesis, can be considered as compositional data with *infinite* number of components, i.e., when $D \to \infty$. And therefore as their functional counterparts: positive (Borel measurable) functions with the unit integral constraint. But using the same arguments as before, also PDFs are rather scale invariant objects and the unit integral constraint is just a proper (despite very important) representative of the equivalence class of proportional functions. This was recognized in Egozcue et al. (2006) and van den Boogaart et al. (2014) equipped the sample space with the Hilbert space structure and called it *Bayes space* – to honor Bayeasian statistics where proportionality of PDFs is commonly used and to point out that the Bayes theorem is essentially the shift from prior to posterior distributions by the likelihood from the Bayes space viewpoint (van den Boogaart et al., 2010). The concepts used in compositional data analysis have their infinitesimal counterparts including the possibility of representing PDFs in the standard L^2 space, where popular methods of functional data analysis can be applied (Ramsay and Silverman, 2005; Kokoszka and Reimherr, 2015). But the more, similar to the Bayes theorem, the general setting of Bayes spaces enables to bring many well-known concepts down to simple algebra and it allows to elegantly define new methods. We will demonstrate all of this in the following sections.

2 Bayes spaces as a general framework for representation of relative data

Bayes spaces are designed to provide a geometric representation for relative data – multivariate compositional data (Aitchison, 1986), compositional tables and cubes (Egozcue et al., 2015; Fačevicová et al., 2022), univariate and multivariate density functions (Egozcue et al., 2006; Genest et al., 2023). – characterised by the property of scale invariance (van den Boogaart et al., 2014). This property states that, given either a finite or infinite domain Ω and a positive real multiple c, two proportional positive functions $f(x)$ and $g(x)$ (i.e. such that $g(x) = cf(x)$, for $c > 0$) carry essentially the same relative information (van den Boogaart et al., 2014). This also follows the common strategy in Bayesian statistics, where multiplicative factors are typically omitted from computations, as they are not essential to the definition of the distributions at hand. Note that the scale invariance of a (discrete or continuous) density f is a direct consequence of the same property of the associated measure μ, i.e. the σ-finite measure μ, such that $f = d\mu/d\mathsf{P}$ for a reference measure P. In this context we refer to the so-called $\mathcal{B}$-equivalence of measures (and densities): two measures μ and ν are $\mathcal{B}$-equivalent if they are proportional, i.e. there exists a positive real multiple c such that $\nu(\mathsf{A}) = c \cdot \mu(\mathsf{A})$ for any $\mathsf{A} \in \mathcal{A}$, $\mathcal{A}$ being a sigma-algebra on Ω.

2.1 Bayes Hilbert spaces

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. P, i.e.,

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

From a practical point of view, the reference measure P plays an important role in the whole concept, as thoroughly investigated in Talská et al. (2020). The choice of the reference measure determines a weighting of the domain Ω of the composition or PDF (i.e. of the distribution in the discrete or continuous case), which can be used to give more relevance to certain regions of Ω when performing multivariate statistics or functional data analysis (FDA), according to the purpose of the analysis

(van den Boogaart et al., 2014; Egozcue and Pawlowsky-Glahn, 2016; Talská et al., 2020). However, even if this is not necessarily of primary interest, the choice of the reference measure makes it possible to cover all the above cases of relative data within the Bayes space framework. Furthermore, to change the reference measure from λ (typically the Lebesgue or uniform measure) to a measure P with strictly positive λ density $p = d\mathsf{P}/d\lambda$, the well-known chain rule can be used. For a generic measure μ we have

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

As mentioned above, the Bayes space framework covers the usual (unweighted) case of D-part compositional data for $\Omega = \{1, \ldots, D\}$ and by taking P as the counting measure, that is, for $x = 1, \ldots, D$, $\mathrm{P}(\{\mathrm{x}\}) = 1$. The set of vectors of $\mathbb{R}^D$ with positive components are densities of measures on Ω, and they are $\mathcal{B}$-equivalent if they have proportional components. In this case, $\mathcal{B}^2(\mathsf{P})$ is a $(D-1)$-dimensional Euclidean space (Pawlowsky-Glahn and Egozcue, 2001; Billheimer et al., 2001) and $f(x) = (f(1), \ldots, f(D)) \equiv (x_1, \ldots, x_D)$. The special case of compositional data analysis is addressed in many publications, e.g. van den Boogaart and Tolosana-Delgado (2013); Pawlowsky-Glahn et al. (2015); Filzmoser et al. (2018); Greenacre (2018), which are more or less grounded in the underlying geometric properties. Similarly, Bayes spaces can be defined for the case where the domain Ω is a Cartesian product of two domains Ω_X and Ω_Y, i.e. $\Omega = \Omega_X \times \Omega_Y$. In this case the reference measure P can be decomposed as a product measure $\mathsf{P} = \mathsf{P}_X \times \mathsf{P}_Y$ and the Hilbert space structure of the Bayes space $\mathcal{B}^2(\mathsf{P})$ can be constructed accordingly (Hron et al., 2023; Genest et al., 2023). This covers the cases of compositional tables and bivariate densities, specifically $f(x) = [f(i,j)] \equiv [x_{ij}]$ for $\Omega_X = \{1, \ldots, i, \ldots, I\}$, $\Omega_Y = \{1, \ldots, j, \ldots, J\}$ in the former case (Egozcue et al., 2015; Fačevicová et al., 2018) and $f(x) \equiv f(x,y)$ for $\Omega_X, \Omega_Y \subset \mathbb{R}$ in the latter case (Hron et al., 2023; Genest et al., 2023); both can be extended to the multivariate setting (Fačevicová et al., 2022; Genest et al., 2023).

The Bayes space is built with operations of *perturbation* and *powering* which can be defined for any two densities f, g with respect to P and a real constant α as

$$(f \oplus g)(x) =_{\mathcal{B}^2(\mathsf{P})} f(x) \cdot g(x) \quad \text{and} \quad (\alpha \odot f)(x) =_{\mathcal{B}^2(\mathsf{P})} f(x)^{\alpha}, \tag{2}$$

respectively. The lower index in $=_{\mathcal{B}^2(\mathsf{P})}$ means that the right hand side of the equations can be arbitrarily rescaled without altering the relative information that the resulting density in $\mathcal{B}^2(\mathsf{P})$ contains. The Hilbert space structure is completed by defining the Bayes inner product,

$$\langle f, g\rangle_{\mathcal{B}^2(\mathsf{P})} = \frac{1}{2\mathsf{P}(\Omega)} \int_\Omega \int_\Omega \ln \frac{f(x)}{f(s)} \ln \frac{g(x)}{g(s)}\, d\mathsf{P}(x)\, d\mathsf{P}(s), \tag{3}$$

which implies in the usual way also the norm and the distance,

$$||f||_{\mathcal{B}^2(\mathsf{P})} = \sqrt{\langle f, f\rangle_{\mathcal{B}^2(\mathsf{P})}}, \quad d_{\mathcal{B}^2(\mathsf{P})}(f, g) = ||f \ominus g||_{\mathcal{B}^2(\mathsf{P})}, \tag{4}$$

where $f \ominus g = f \oplus [(-1) \odot g]$ is the perturbation-subtraction of densities. Note that in case of compositional data or compositional tables we refer for (2)-(4) to the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001) which was developed historically before the first concepts of Bayes spaces were introduced (Egozcue et al., 2006). While the scale of the reference measure P does not have any impact for the operations of perturbation and powering, it does influence the inner product because changing the scale corresponds to shrinkage (or expansion) of the Bayes space (for details, see Talská et al. (2020)).

The usual strategy when dealing with the Bayes spaces (van den Boogaart et al., 2014; Hron et al., 2016; Talská et al., 2018, 2021) is not to process densities directly in the original space but to map them into the standard L^2 (Euclidean) space where most of the widely-used methods of functional (multivariate) data analysis can be employed. The *clr transformation* of a density $f(x) \in \mathcal{B}^2(\mathsf{P})$ is a real function $f^c : \Omega \to \mathbb{R}$, $f^c \in L^2_0(\mathsf{P})$, defined as

$$f^c(x) = \mathrm{clr}(f)(x) = \ln f(x) - \frac{1}{\mathsf{P}(\Omega)} \int_\Omega \ln f(x)\, d\mathsf{P}. \tag{5}$$

Similar as for perturbation and powering, the scale of P does not play any role in (5), too. On the other hand, one should note that the resulting function f^c is expressed with respect to reference P. As a consequence, using any measure other than the Lebesgue (or uniform) λ leads to clr-transformations defined over a weighted space and a further "unweighting" step is needed (Egozcue and Pawlowsky-Glahn, 2016; Talská et al., 2020). Moreover, one should also take into account the zero-integral constraint of clr transformed densities, i.e.,

$$\int_\Omega f^c(x)\, d\mathsf{P} = 0. \tag{6}$$

Here $L_0^2(\mathsf{P})$ denotes the subspace of the $L^2(\mathsf{P})$ space of real functions having zero integral; in particular, one clearly has that $f^c(x) \in L_0^2(\mathsf{P})$. In the functional case, this constraint usually does not represent any serious obstacle for the application of FDA methods – especially if a proper spline representation of the PDFs, compatible with the concept of Bayes spaces, is used (Machalová et al., 2016, 2021; Hron et al., 2023) which forms a cornerstone in a large number of computational methods for FDA.

A prominent case to see the effect of density data analysis in Bayes spaces is that of functional principal component analysis of PDFs, which we refer to as *simplicial functional principal component analysis (SFPCA)* (Hron et al., 2016), recently extended to incorporate also measurement process errors (Pavlů et al., 2024). The aim of SFPCA is analogous to that of PCA on multivariate data: to capture the main modes of variability in the data by a small number K of linear combinations of the original variables. Let λ be the commonly used Lebesgue reference measure and all PDFs are defined on the same domain $I = \langle a, b\rangle$. Then, for $X_1, ..., X_N$ being a (centred) sample in $\mathcal{B}^2(\lambda)$, i.e., we performed perturbation-subtraction by $\overline{X} = \frac{1}{N} \odot \bigoplus_{i=1}^{N} X_i$, SFPCA looks firstly for the main mode of variability. This means, for the element ζ_1 in $\mathcal{B}^2(\lambda)$ – called first simplicial functional principal component (SFPC)– maximizing over $\zeta \in \mathcal{B}^2(\lambda)$,

$$\frac{1}{N} \sum_{i=1}^{N} \langle X_i, \zeta \rangle_{\mathcal{B}^2(\lambda)}{}^2 \text{ subject to } \|\zeta\|_{\mathcal{B}^2(\lambda)} = 1;\ \langle \zeta_j, \zeta_k \rangle_{\mathcal{B}^2(\lambda)} = 0,\ k < j.$$

The remaining SFPCs, $\{\zeta_j\}_{j \geq 2}$, capture the remaining modes of variability subject to be mutually orthogonal, and are thus obtained by solving problem the previous maximization problem with the additional orthogonality constraint $\langle \zeta_k, \zeta \rangle_{\mathcal{B}^2(\lambda)} = 0, k < j$. The output are *eigenfunctions* ζ_j of the sample covariance operator $V : \mathcal{B}^2(\lambda) \to \mathcal{B}^2(\lambda)$, acting on $x \in \mathcal{B}^2(\lambda)$ as

$$Vx = \frac{1}{N} \bigoplus_{i=1}^{N} \langle X_i, x \rangle_{\mathcal{B}^2(\lambda)} \odot X_i,$$

also called harmonics (interpreted in terms of the original data), and *scores* (coefficients, representing data structure of the original observations), so that finally

$$X_i \approx \sum_{k=1}^{K} \langle X_i, \zeta_k \rangle_{\mathcal{B}^2(\lambda)} \odot \zeta_k.$$

The j-th SFPC ζ_j and the associated scores $\Psi_{ij} = \langle X_i, \zeta_j \rangle_{\mathcal{B}^2(\lambda)}$, $i = 1, ..., N$, are obtained by solving the eigenvalue equation

$$V\zeta_j = \rho_j \odot \zeta_j; \tag{7}$$

ρ_j denotes the j-th eigenvalue, with $\rho_1 \geq \rho_2 \geq ...$. For each j, the term $\rho_j / \sum_j \rho_j$ is associated with the proportion of total variability explained by the SFPC ζ_j. The eigenvalue equation is solved using basis expansion of each datum X_i, $i = 1, ..., N$ using K known basis functions $\phi_1, ..., \phi_K$:

$$X_i(\cdot) = \bigoplus_{k=1}^{K} c_{ik} \odot \phi_k(\cdot),$$

where $c_{ik} = \langle X_i, \phi_k \rangle_{\mathcal{B}^2(\lambda)}$, $k = 1, ..., K$. Commonly, *compositional splines* (Machalová et al., 2021) are used for this purpose. To perform SFPCA exploiting the efficient routines available in L^2 space (i.e., avoid computations in Bayes spaces), the clr transformation (5) can be used which maps the operations of perturbation and powering and the Bayes inner product to the usual addition of two real functions and multiplication of a function by a scalar, and the standard L^2 inner product, specifically

- $\mathrm{clr}(f \oplus g)(t) = f^c(t) + g^c(t), \quad \mathrm{clr}(\alpha \odot f)(t) = \alpha \cdot f^c(t), \quad t \in I,$
- $\langle f, g \rangle_{\mathcal{B}^2(I)} = \langle \mathrm{clr}(f), \mathrm{clr}(g) \rangle_{L^2(I)}.$

However, analysing PDFs in Bayes spaces typically adds value beyond representation in a meaningful sample space. In the context of dimension reduction, an interesting property can be exploited for PDFs belonging to the extended exponential family. Recall that a *k-parametric extended exponential family* on Ω, $Exp_{\mathcal{B}^2(I)}(g, \boldsymbol{T}, \boldsymbol{\vartheta})$ is a collection of densities

$$f(t, \boldsymbol{\alpha}) =_{\mathcal{B}^2(I)} g(t) \cdot \exp\left\{ \sum_{j=1}^{k} \vartheta_j(\boldsymbol{\alpha}) T_j(t) \right\}, \quad t \in \Omega,$$

where $\boldsymbol{\alpha}$ denotes the k-dimensional vector of parameters in a k-dimensional parameter space A, while functions $g : \Omega \to \mathbb{R}$, $\vartheta_j : A \to \mathbb{R}$ and $T_j : \Omega \to \mathbb{R}$, $j = 1, ..., k$, are Borel-measurable. An extended exponential family on Ω is a finite dimensional

affine subspace of the Bayes space $\mathcal{B}^2(I)$ (van den Boogaart et al., 2010). A PDF in $Exp_{\mathcal{B}(I)}(g, \boldsymbol{T}, \boldsymbol{\vartheta})$ can then be expressed as a linear combination in $\mathcal{B}^2(I)$,

$$f(t, \boldsymbol{\alpha}) =_{\mathcal{B}^2(I)} g(t) \oplus \bigoplus_{j=1}^{k} \left[\vartheta_j(\boldsymbol{\alpha}) \odot \exp\{T_j(t)\}\right], \quad t \in \Omega,$$

or equivalently in the clr space as

$$f^c(t, \boldsymbol{\alpha}) = \mathrm{clr}(g(t)) + \sum_{j=1}^{k} \left[\vartheta_j(\boldsymbol{\alpha}) \cdot \mathrm{clr}(\exp\{T_j(t)\})\right], \quad t \in \Omega.$$

For $k_0 \leq k$ uncertain parameters, the SFPCA thus estimates an orthonormal basis of the corresponding k-dimensional affine space in $\mathcal{B}^2(I)$, which is associated to $k_0 \leq k$ non-zero eigenvalues. This will be illustrated with an example from Hron et al. (2016).

Example 1 (truncated normal) *We consider normal densities,* $\mu = 0, \sigma_i = \exp(-1 + (i-1)/10)$*,* $i = 1, ..., 21$*,* $I = [-5, 5]$

$$f(t; \sigma_i) =_{\mathcal{B}(\lambda)^2} \exp\left\{-\frac{t^2}{2\sigma_i^2}\right\}, \quad t \in I, \tag{8}$$

or in the clr space

$$f^c(t; \sigma_i) = -\frac{t^2}{2\sigma_i^2} + \frac{25}{6\sigma_i^2}, \quad t \in I,$$

see Figure 1. A normal density $N(0, \sigma^2)$ *restricted on* Ω *belongs to a 1-parametric extended exponential family, with* $\alpha = \sigma$*,* $\vartheta_1(\alpha) = 1/\sigma^2$*, and* $T_1(t) = -t^2$*. Figure 2 reports the results of SFPCA. The first SFPC –displayed in Figure 2c– captures the entire variability of the dataset and is precisely interpreted in terms of mass concentration. Indeed, the positive scores along the first SFPC are associated with the highest standard deviations of the set of normal PDFs and vice versa (Figure 2b; here the indices* $i = 1, ..., 21$ *refer to the standard deviation* σ_i *of the corresponding density). Such an interpretation can be easily derived from the plot of the mean density perturbed by* $\oplus/\ominus$ *the first SFPC weighted according to the corresponding standard deviation – i.e.,* $\sqrt{\rho_1}$*,* ρ_1 *appearing in (7) –, that is depicted in Figure 2e.*

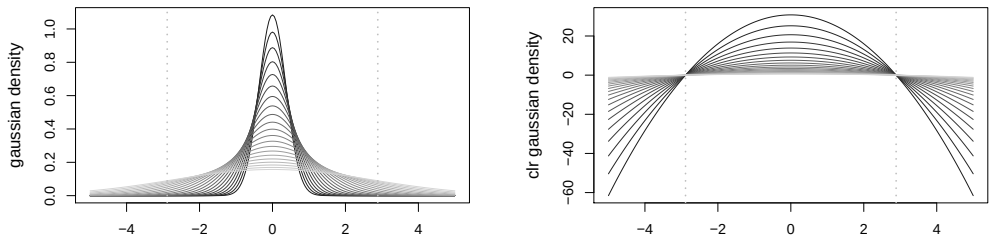

Figure 1: Simulated truncated normal densities in the original space (left) and in the clr space (right).

2.2 Multivariate Bayes spaces

When dealing with multivariate PDFs, the natural goal is to filter out interactions from the independent part, commonly understood as the product of the marginals. A classical result is Sklar's theorem (Sklar, 1959), which states that for a PDF $f(x_1, \ldots, x_d)$ of a random vector $(X_1, \ldots, X_d)$ in the domain $I = [a_1, b_1] \times \ldots, [a_d, b_d]$, marginal PDFs $f_1(x_1), \ldots, f_d(x_d)$ and marginal distribution functions $F_1(x_1), \ldots, F_d(x_d)$ we can write

$$f(x_1, \ldots, x_d) = f_1(x_1) \cdot \ldots \cdot f_d(x_d) \cdot c(F_1(x_1), \ldots, F_d(x_d)), \tag{9}$$

where c is the density of a copula

$$C(u_1, u_2, \ldots, u_d) = \mathsf{P}(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d)$$

for $U_i = F_i(X_i)$ are the uniform random variables corresponding to the marginals $F_i(x_i)$ of X_i, $i = 1, \ldots, d$. While the standard (arithmetic) marginal PDFs occur in (9), further decomposition of c in the original domain is not easily possible. Moreover, it is not specified whether there is a geometric relation between the marginal PDFs $f_1, \ldots, f_d$ and the "interaction component" c.

All this can be achieved by embedding f in the Bayes space setting. Let $(\Omega, \mathcal{A})$ be a d-dimensional product space and let λ be a finite product reference measure. Specifically, suppose that for $i \in D$, $(\Omega_i, \mathcal{A}_i)$ is a measurable space and λ_i is a finite, positive, real-valued measure on it. Set

$$\Omega = \Omega_1 \times \cdots \times \Omega_d, \quad \mathcal{A} = \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_d, \quad \lambda = \lambda_1 \otimes \cdots \otimes \lambda_d.$$

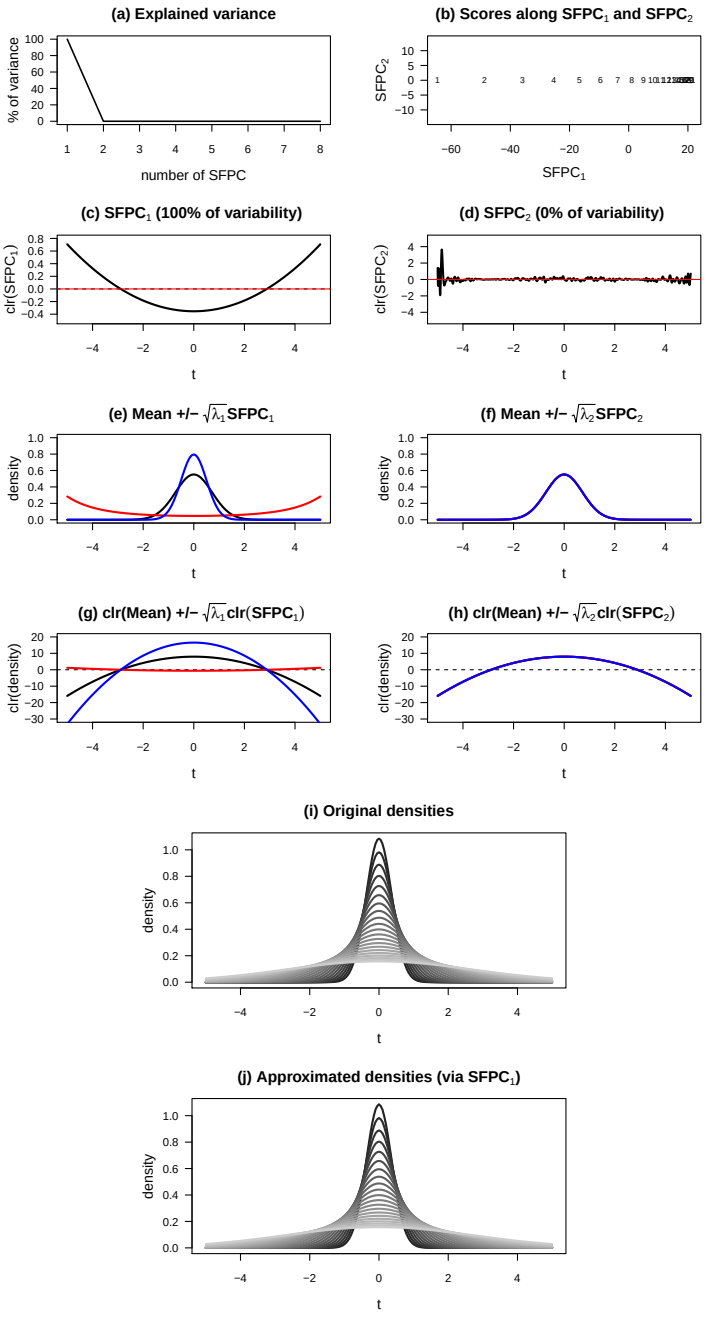

Figure 2: SFPCA of Gaussian densities with $\mu = 0$ and $\sigma_i = \exp(-1 + (i-1)/10)$ for $i = 1, \ldots, 21$.

For arbitrary non-empty $I \subseteq D$, let

$$\Omega_I = \times_{i\in I}\Omega_i, \quad \mathcal{A}_I = \bigotimes_{i\in I}\mathcal{A}_i, \quad \lambda_I = \bigotimes_{i\in I}\lambda_i.$$

The need to have the product space is the only restriction we need to take into account. For the reference measure, the Lebesgue measure is the first choice for FDA, together with the usual case of the bounded domain, but more general settings can also be considered. The Hilbert space structure of multivariate Bayes spaces can be defined analogously, we refer to Genest et al. (2023) for further details.

As stated at the beginning of this section, the consideration of classical arithmetic marginals for the decomposition of the multivariate density f has its limitations. We will instead introduce the concept of geometric marginals, which is a *generalisation* of the former and allows a much deeper insight into the dependence structure of PDFs.

Let's consider $\mathcal{B}^2_I(\lambda)$ to be a subspace of $\mathcal{B}^2(\lambda)$ with PDFs of arguments with indices in I only. For $I \subsetneq D$, the *I-th geometric marginal* is

$$f_{I,g} = \exp\left\{\frac{1}{\lambda_{D\setminus I}(\Omega_{D\setminus I})}\int_{\Omega_{D\setminus I}} \ln(f)d\lambda_{D\setminus I}\right\}. \tag{10}$$

and its clr transformation

$$\mathrm{clr}(f_{I,g}) = \frac{1}{\lambda_{D\setminus I}(\Omega_{D\setminus I})}\int_{\Omega_{D\setminus I}} \mathrm{clr}(f)d\lambda_{D\setminus I}.$$

The logarithm of f in $f_{I,g}$ can be replaced by its clr transformation due to properties of the exponential (and the proportionality of densities in the Bayes space framework). Nevertheless, the form in which the geometric marginal is defined in (11) clearly indicates its origin as a generalisation of the geometric mean of positive data. As a special case, for $I = \{i\}$, $i = 1, \ldots, d$ the *i-th geometric marginal* is

$$f_{i,g} = \exp\left\{\frac{1}{\lambda_{D\setminus\{i\}}(\Omega_{D\setminus\{i\}})}\int_{\Omega_{D\setminus\{i\}}} \ln(f)d\lambda_{D\setminus\{i\}}\right\}. \tag{11}$$

The I-th geometric marginal is the unique *orthogonal projection* of $f(\equiv f_\mu)$ onto $\mathcal{B}^2_I(\lambda)$, a property which has no counterpart in the concept of (classical) arithmetic marginals. And this is also a clear interpretive advance of the geometric marginals.

In the case of true independence, when f is a product of (arithmetic) marginals, the i-th geometric and the corresponding arithmetic marginals coincide. Consequently, from this point of view, arithmetic marginals can be considered as their geometric counterparts *under the assumption of independence*. Geometric marginals also contain (univariate) information from the dependence structure. Consequently, the i-th geometric marginals of f and of f^*, where $D^* \subset D$, are generally different. There is also another point to consider. Looking at the decomposition (9), the i-th geometric marginal can be formulated as

$$f_{i,g} = \exp\left\{ \frac{1}{\lambda_{D\setminus\{i\}}(\Omega_{D\setminus\{i\}})} \int_{\Omega_{D\setminus\{i\}}} \Big[\ln\{c(F_1,\ldots,F_d)\} + \sum_{j=1}^{d} \ln(f_j) \Big] d\lambda_{D\setminus\{i\}} \right\}.$$

Since for any distinct $i, j \in D$ the integral of $\ln(f_j)$ over $\Omega_{D\setminus\{i\}}$ is a constant, we find that $f_{i,g} \propto f_i g_i$, where

$$g_i = \exp\left\{ \frac{1}{\lambda_{D\setminus\{i\}}(\Omega_{D\setminus\{i\}})} \int_{\Omega_{D\setminus\{i\}}} \ln\{c(F_1,\ldots,F_d)\} d\lambda_{D\setminus\{i\}} \right\} \tag{12}$$

depends on both the copula and the marginals $F_1, \ldots, F_d$. Consequently, if there are parameters specifically related to some (arithmetic) marginal, they will propagate through the dependence structure to other geometric marginals. However, this can be seen as an inherent property of the dependence structure rather than a deficiency of the geometric marginals: *i-th geometric marginals are those that capture all the univariate information about i-th variable from the multivariate structure of f*. We illustrate this with two examples for $d = 2$. The first is from Hron et al. (2023):

Example 2 (truncated bivariate normal) *We consider a zero-mean bivariate Gaussian density $\mathcal{N}_2(\mu, \Sigma)$ with respect to the (product) Lebesgue measure $\lambda[I] = \lambda[I_1] \times \lambda[I_2]$, truncated on a rectangular domain $I = I_1 \times I_2 \subset \mathbb{R}^2$, with $I_1 = I_2 = [-T, T]$, $T = 5$. The PDF is defined, for $\mathbf{x} = (x, y) \in I$, as*

$$f(x,y) =_{\mathcal{B}^2(\lambda)} \exp\left\{ \mathbf{x}^\top \Sigma^{-1} \mathbf{x} \right\} = \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{x^2}{\sigma_1^2} - 2\rho \frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right] \right\}, \tag{13}$$

where $\sigma_i^2 = \Sigma_{ii}$ and $\rho \in [-1, 1]$ is the correlation coefficient. Clearly, for $\rho = 0$ the case of independence is achieved. The clr transformation of f yields

$$\mathrm{clr}(f)(x,y) = -\frac{1}{2(1-\rho^2)} \left[\frac{x^2}{\sigma_1^2} - 2\rho \frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right] + \frac{T^2}{6(1-\rho^2)} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right). \tag{14}$$

The X-th geometric marginal is derived from (14) as

$$\mathrm{clr}(f_X)(x) = -\frac{1}{2(1-\rho^2)} \cdot \frac{x^2}{\sigma_1^2} + \frac{T^2}{6(1-\rho^2)} \cdot \frac{1}{\sigma_1^2}, \quad x \in I_1$$

and eventually back-transformed to the original space as

$$f_X(x) =_{\mathcal{B}^2(\lambda_X)} \exp\left[-\frac{1}{2(1-\rho^2)}\frac{x^2}{\sigma_1^2}\right]$$

(analogously for the Y-th marginal). Obviously, the parameter ρ for $\rho \neq 0$ propagates to the respective geometric marginals, which are again truncated normal PDFs (see Figure 3), and reduces their variance. In turn, it is interesting to analyse what can be observed from the perturbation difference $f_{X,g} \ominus f_X$, which results in

$$f_{X,g} \ominus f_X(x) =_{\mathcal{B}^2(\lambda_X)} \exp\left[-\frac{\rho^2}{2(1-\rho^2)}\frac{x^2}{\sigma_1^2}\right]$$

and indicates how much univariate information about X interferes with the dependence structure (see Figure 4). Starting from the uniform PDF for $\rho = 0$ as expected, which plays the role of the origin for the Lebesgue reference, the effect of increasing ρ is more pronounced. On the other hand, geometric marginals are not affected by the parameter σ_2 of the marginal distribution of Y, unlike the situation in the next example from Genest et al. (2023).

Example 3 (bivariate beta) *Consider the following three-parameter bivariate beta distribution density. For arbitrary $\alpha_0, \alpha_1, \alpha_2 \in (0, \infty)$ and all $x_1, x_2 \in (0, 1)$, let*

$$f_\mu(x_1, x_2) = \frac{1}{B(\alpha_0, \alpha_1, \alpha_2)} \frac{x_1^{\alpha_1 - 1}(1-x_1)^{\alpha_0+\alpha_2-1} x_2^{\alpha_2-1}(1-x_2)^{\alpha_0+\alpha_1-1}}{(1-x_1x_2)^{\alpha_0+\alpha_1+\alpha_2}},$$

where $B(\alpha_0, \alpha_1, \alpha_2) = \Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_0 + \alpha_1 + \alpha_2)$ is the generalised beta function. The univariate marginals are beta distributions with parameters (α_1, α_0) and (α_2, α_0), respectively, which again correspond to geometric marginals under the assumption of independence, i.e. for $\alpha_0 + \alpha_1 + \alpha_2 \to 0$. This distribution was analyzed in detail from the Bayes spaces perspective in Genest et al. (2023). Here we just note that The corresponding geometric marginals are then given, for all $x_1, x_2 \in (0, 1)$ and $j \in \{1, 2\}$, by

$$\begin{aligned}\mathrm{clr}(f_{\mu,j})(x_j) = (\alpha_j - 1)\ln(x_j) + (\alpha_0 + \alpha_j - 1)\ln(1 - x_j) \\ + (\alpha_0 + \alpha_1 + \alpha_2)\left\{\frac{\pi^2}{6} + \frac{(1-x_1)\ln(1-x_1)}{x_1}\right\} - 2,\end{aligned}$$

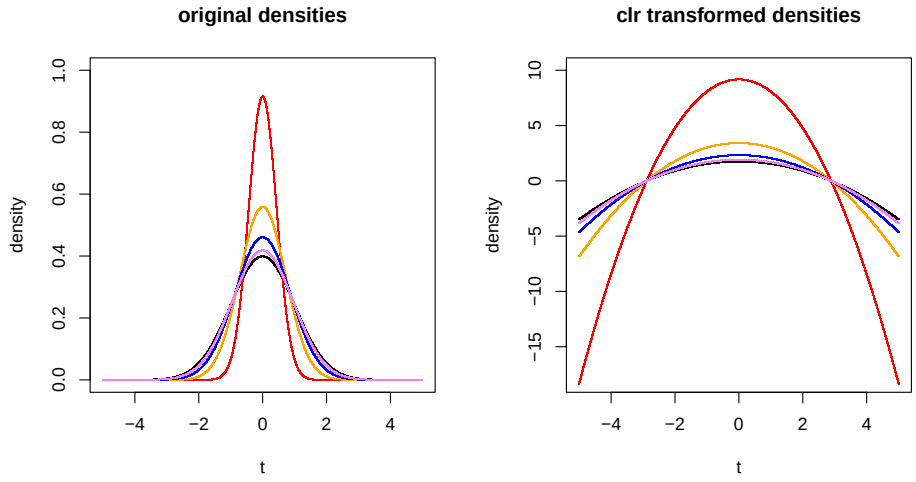

Figure 3: Geometric marginals of X in the truncated bivariate normal PDF with $\sigma_1 = 1$. Colours stand for values of the parameter ρ: $\rho = 0$ (black), $\rho = 0.3$ (violet), $\rho = 0.5$ (blue), $\rho = 0.7$ (orange) and $\rho = 0.9$ (red).

i.e. they actually contain parameters of the latter marginal distribution.

The concept of geometric marginals has many geometric and probabilistic implications which have no counterparts in the case of arithmetic marginals. In Genest et al. (2023) the Hoeffding-Sobol decomposition was used to derive the following property. For any $f \in \mathcal{B}^2(\lambda)$,

$$f = f_{\text{ind}} \oplus \bigoplus_{I \subseteq D, |I| \geq 2} f_{I,\text{int}}$$

where the so-called independence and interaction parts are respectively given by

$$f_{\text{ind}} = \bigoplus_{i=1}^{d} f_i, \quad f_{I,\text{int}} = \bigoplus_{J \subseteq I, J \neq \varnothing} \{(-1)^{|I \setminus J|}\} \odot f_J.$$

and all components of the decomposition are orthogonal to each other. In other words, it is possible to extract the "independent" case of the product of univariate marginals (i-th geometric marginals, $i = 1, \ldots, d$), which are also all mutually

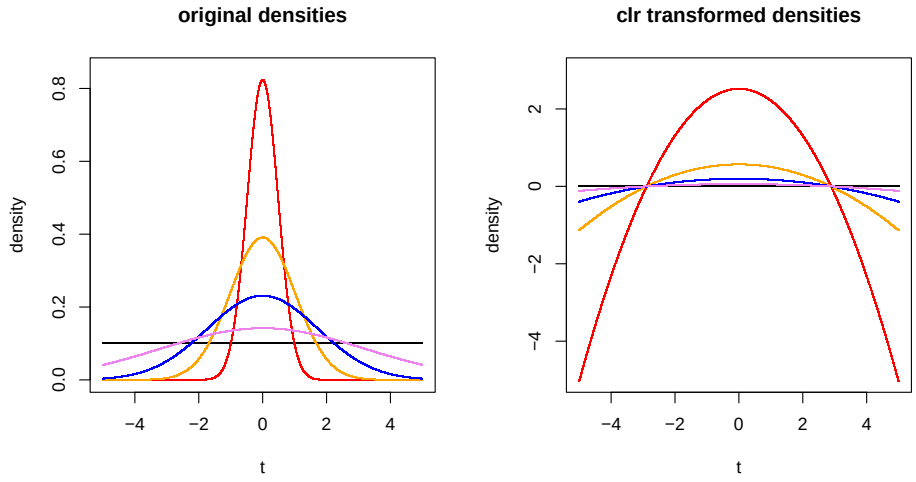

Figure 4: Difference between geometric and arithmetic marginals of X in the truncated bivariate normal PDF with $\sigma_1 = 1$. Colours stand for values of the parameter ρ: $\rho = 0$ (black), $\rho = 0.3$ (violet), $\rho = 0.5$ (blue), $\rho = 0.7$ (orange) and $\rho = 0.9$ (red).

orthogonal, and then use I-th geometric marginals to specify from bivariate interactions up to a PDF representing mutual interactions of all variables. Obviously, in the truly independent case (in the probabilistic sense), the independent part becomes a product of arithmetic marginals and the interaction part becomes a uniform PDF.

One of the most important implications of the orthogonal decomposition of multivariate PDFs is the Pythagorean Theorem,

$$\|f\|^2_{\mathcal{B}^2(\lambda)} = \|f_{\text{ind}}\|^2_{\mathcal{B}^2(\lambda)} + \sum_{I \subseteq D, |I| \geq 2} \|f_{I,\text{int}}\|^2_{\mathcal{B}^2(\lambda)}, \tag{15}$$

where $\|f_{\text{ind}}\|^2_{\mathcal{B}^2(\lambda)} = \|f_1\|^2_{\mathcal{B}^2(\lambda)} + \cdots + \|f_d\|^2_{\mathcal{B}^2(\lambda)}$. Since the Bayes norm can serve as a scalar measure of information (Egozcue and Pawlowsky-Glahn, 2018), (15) can also be interpreted as a decomposition of the information conveyed by the multivariate PDF. Accordingly, the $2^d - 1$ components of the vector

$$R^2_{\mathcal{B}^2(\lambda)}(f) = \left(\frac{\|f_1\|^2_{\mathcal{B}^2(\lambda)}}{\|f\|^2_{\mathcal{B}^2(\lambda)}}, \cdots, \frac{\|f_d\|^2_{\mathcal{B}^2(\lambda)}}{\|f\|^2_{\mathcal{B}^2(\lambda)}}, \frac{\|f_{\{1,2\},\text{int}}\|^2_{\mathcal{B}^2(\lambda)}}{\|f\|^2_{\mathcal{B}^2(\lambda)}}, \cdots, \frac{\|f_{D,\text{int}}\|^2_{\mathcal{B}^2(\lambda)}}{\|f\|^2_{\mathcal{B}^2(\lambda)}} \right) \tag{16}$$

add up to 1. The components of this vector thus provide a measure of the relative contribution of each and every subset of variables in the set D to the overall dependence structure. In the spirit of Egozcue and Pawlowsky-Glahn (2018), the compositional vector (16) could also be regarded as a break-down of the total information contained in the density f_μ. The sum of the terms associated with subsets of size 2 and above constitutes the multivariate analog of the simplicial deviance from Hron et al. (2023). The resulting information composition can be analyzed using methods of compositional data analysis, like those collected in (Filzmoser et al., 2018). Let's illustrate this on example from Matys Grygar et al. (2024):

Example 4 (RKP data) *We consider trivariate densities of (log-transformed) copper (Cu), lead (Pb) and zinc (Zn) soil concentration data in 77 districts of the Czech Republic from the Register of Contaminated Areas (registr kontaminovaných ploch) collected by the Department of Agriculture of the Czech Republic (Podlešáková et al., 1996; Zbíral et al., 2004; Poláková et al., 2011). At least higher hundreds of concentration values were available in each district and smoothed to density data. Districts are characterized by diverse geological origin and anthropogenic contamination, which is however typically homogeneous enough within one district.*

In Figure 5 the result of hierarchical clustering of information compositions in all districts using complete linkage is presented in form of a heatmap. For this purpose the compositions were represented first with the multivariate – counting measure – version of the clr transformation (5). Compositional parts corresponding to the univariate geometric marginals are denoted $f(\mathrm{Cu})$, $f(\mathrm{Pb})$, *and* $f(\mathrm{Zn})$, *to the bivariate interactions* $f(\mathrm{Cu},\mathrm{Pb})$, $f(\mathrm{Cu},\mathrm{Zn})$, *and* $f(\mathrm{Pb},\mathrm{Zn})$, *and to the trivariate interaction* $f(\mathrm{Cu},\mathrm{Pb},\mathrm{Zn})$. *Among other patterns, which are in detail described in Matys Grygar et al. (2024), it is interesting to observe that districts collected in cluster C have all the same source of contamination (agricultural spraying due to intensive land use). They are characterized by simultaneous heterogeneities in uni- and bivariate densities with Cu, which is common also for other pesticide-contaminated districts.*

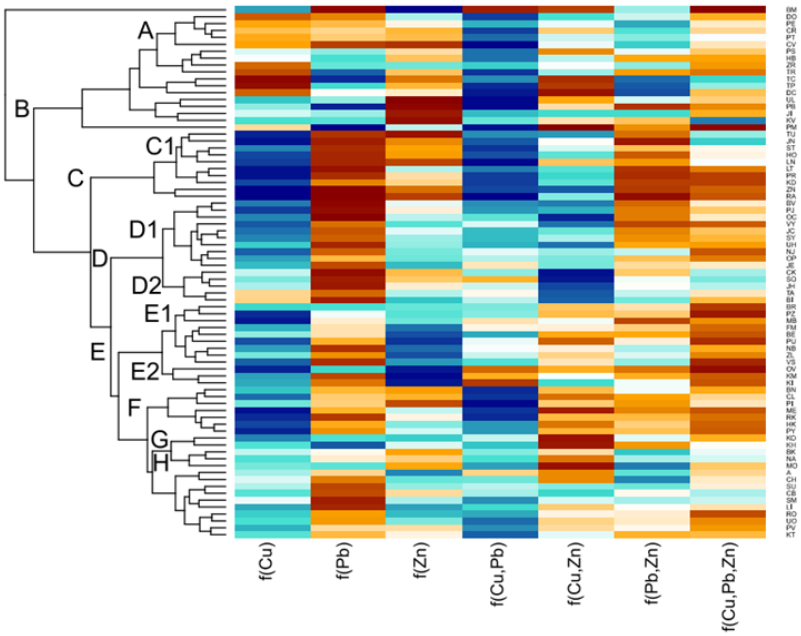

Figure 5: Hierarchical clustering of districts according to their respective clr transformed information compositions. Specific values of norms indicate heterogeneity of the respective distributions: low norms (in blue hues) correspond to high heterogeneities, while high norms (in red hues) evidence narrow distributions.

3 Goals of the thesis

Bayes spaces provide a general framework for embedding discrete and continuous distributions, as well as measures. The thesis aims to present the contributions developed by the Candidate, in collaboration with his PhD students and other colleagues. These contributions have aimed to

1. build a solid theoretical ground for Bayes spaces,

2. demonstrate the potential of Bayes spaces in concrete popular methods of functional data analysis, adapted to the analysis of samples of probability density functions.

In fact, the latter goal was of primary interest at the outset. The motivation was to follow up the initial application of Bayes spaces to population pyramids (Delicado, 2011) in the context of dimension reduction and to show that their potential for adapting methods for statistical processing of PDFs is much broader, but also that the benefits of using Bayes spaces go beyond the possibility of representing PDFs in the usual L^2 space. Clearly, these benefits are primarily related to the scale invariance of densities, which sheds new light on the proportionality of PDFs known from Bayesian statistics and is the key to understanding that the variability of densities is inherently contained in their small function values. In the context of dimension reduction, but also for the representation of PDFs in general, it is also crucial to consider that distributions from the exponential family form a finite dimensional subspace of the Bayes space.

Naturally, the development of methods for statistical processing of PDFs soon led to theoretical challenges related to Bayes spaces. To be able to weight the domain of PDFs by a proper choice of the reference measure, it was necessary to clarify the role of its scale, because there was an inconsistency between the foundational works of Egozcue et al. (2006) and van den Boogaart et al. (2014). This inconsistency led to problems with the continuous counterpart of the principle of subcompositional dominance known from compositional data analysis (Pawlowsky-Glahn et al., 2015); resolving this opened up the possibility of analysing PDFs with a general reference measure using standard methods of FDA. The next step was the development of Bayes spaces for multivariate PDFs, where the key point was to redefine the role

of the marginals and to reveal the role played by the scale of the reference measure. Here Egozcue et al. (2015) and Fačevicová et al. (2022) were instrumental in understanding the structure of multivariate Bayes spaces.

4 Structure of the thesis

The thesis consists of six papers which can be divided into two main blocks according to the goals of the thesis. In the first block, three papers are presented that demonstrate the potential of Bayes spaces for functional data analysis of samples of probability density functions, also known as density data analysis. The second block contains three theoretical papers that develop the Bayes space methodology itself, but also provide impetus for concrete applications. For each paper, there is also a concrete specification of the Candidate's contribution, which in some papers reflects his role as supervisor of a PhD student.

These papers can be considered as a concise selection of the Candidate's scientific output, although they do not represent a complete list of his work on Bayes spaces. Some other closely related papers not included in the thesis are therefore listed in Section 2.

4.1 Density data analysis using Bayes spaces

- Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P. (2016) *Simplicial principal component analysis for density functions in Bayes spaces.* Computational Statistics and Data Analysis 94, 330–350.

The aim of the paper is to build up a concise methodology for functional principal component analysis of densities. A simplicial functional principal component analysis (SFPCA) is proposed, based on the geometry of the Bayes space of functional compositions. SFPCA is performed by exploiting the centred log-ratio transform, an isometric isomorphism between the Bayes space and the L^2 space which enables one to resort to standard functional data analysis tools. The advantages of the proposed approach with respect to existing techniques are demonstrated using simulated data and a real-world example of population pyramids in Upper Austria.

Contribution of the Candidate:

- Development of the SFPCA model.

- Design and evaluation of the simulation study.

- Interpretation of the results of the empirical study.

- Talská, R., Hron, K., Matys Grygar, T. (2021) *Compositional scalar-on-function regression with application to sediment particle size distributions.* Mathematical Geosciences 53, 1667–1695.

The chemical composition of sediments is controlled predominantly by the sediment grain size, and thus evaluating their relationship is an important task in sedimentary geochemistry. The grain size is characterized by the respective particle size distribution, which can be expressed as a probability density function. Because of the relative character of densities, the Bayes space methodology was employed to build a functional regression model between a real response and a density function as a covariate, here the chemical composition and the particle size density. For practical computations, density functions were expressed in the standard L^2 space using the centred logratio transformation and spline approximation of the input discretized densities was utilized by respecting the induced zero-integral constraint. After a concise simulation study, supporting the relevance of the proposed regression model, the new methodology was applied to examine the relationship between sediment grain size and geochemical composition, with samples being obtained in the Czech Republic in the Skalka Reservoir and in the Ohře River floodplain upstream of the reservoir, to reveal proper grain size proxies. The Al/Si and Zr/Rb logratios in the sediments that were studied showed grain-size control, which makes them suitable for this purpose.

Contribution of the Candidate:

- Development of the compositional scalar-on-function regression model.

- Design and evaluation of the simulation study.

- Interpretation of the results of the empirical study.

- Pavlů, I., Menafoglio, A., Bongiorno, E.G., Hron, K. (2023) *Classification of probability density functions in the framework of Bayes spaces: methods and applications.* Statistics and Operations Research Transactions 47, 295–322.

In this paper the process of supervised classification when the data set consists of probability density functions is studied. Due to the relative information contained in densities, it is necessary to convert the functional data analysis methods into an appropriate framework, here represented by the Bayes spaces. This work develops Bayes space counterparts to a set of commonly used functional methods with a focus on classification. Hereby, a clear guideline is provided on how some popular classification approaches can be adapted for the case of densities, and that in the classification context it is also quite straightforward. Comparison of the methods is based on simulation studies and real-world applications, reflecting their respective strengths and weaknesses.

Contribution of the Candidate:

- Development of the Bayes space formulation of the classification models.

- Design and evaluation of the simulation study.

- Design of the empirical studies, interpretation of their results.

4.2 Methodological contributions to Bayes spaces

- Talská, R., Menafoglio, A., Hron, K., Egozcue, J.J., Palarea-Albaladejo, J. (2020) *Weighting the domain of probability densities in functional data analysis.* Stat 9, e283.

In functional data analysis, some regions of the domain of the functions can be of more interest than others owing to the quality of measurement, relative scale of the domain, or simply some external reason (e.g. interest of stakeholders). Weighting the domain is of interest particularly with probability density functions (PDFs), as derived from distributional data, which often aggregate measurements of different quality or are affected by scale effects. A weighting scheme can be embedded into the underlying sample space of a PDF when it is considered as continuous compositions applying the theory of Bayes spaces. The origin of a Bayes space is determined by a given reference measure, and this can be easily changed through the well-known chain rule. This work provides a formal framework for defining weights through a reference measure, and it is used to develop a weighting scheme on the

bounded domain of distributional data. The impact on statistical analysis is illustrated through an application to functional principal component analysis of income distribution data. Moreover, a novel centred log-ratio transformation is proposed to map a weighted Bayes space into an unweighted L^2 space, enabling to use most tools developed in functional data analysis (e.g. clustering and regression analysis) while accounting for the weighting scheme. The potential of our proposal is shown on a real case study using Italian income data.

Contribution of the Candidate:

- Development of the weighting scheme.

- Design and evaluation of the simulation study.

- Design of the empirical study, interpretation of its results.

- Hron, K., Machalová, J., Menafoglio, A. (2023) *Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation.* Statistical Papers 64, 1629–1667.

A new orthogonal decomposition for bivariate probability densities embedded in Bayes Hilbert spaces is derived. It allows representing a density into independent and interactive parts, the former being built as the product of revised definitions of marginal densities, and the latter capturing the dependence between the two random variables being studied. The developed framework opens new perspectives for dependence modelling (e.g., through copulas), and allows the analysis of datasets of bivariate densities, in a functional data analysis perspective. A spline representation for bivariate densities is also proposed, providing a computational cornerstone for the developed theory.

Contribution of the Candidate:

- Development of the orthogonal decomposition of bivariate densities.

- Design of the empirical study, processing and interpretation of its results.

- Genest, C., Hron, K., Nešlehová, J.G. (2023) *Orthogonal decomposition of multivariate densities in Bayes spaces and relation with their copula-based representation.* Journal of Multivariate Analysis 198, 105228.

Bayes spaces were initially designed to provide a geometric framework for the modeling and analysis of distributional data. In Hron et al. (2023) it was shown that this methodology can be exploited to construct an orthogonal decomposition of a bivariate probability density into an independence and an interaction part. In this paper, new insights into these results are given by reformulating them using Hilbert space theory, and a multivariate extension is developed using a distributional analog of the Hoeffding–Sobol identity. A connection is also made between the resulting decomposition of a multivariate density and its copula-based representation.

Contribution of the Candidate:

- Conceptualization, methodology, validation, writing – original draft, review and editing (CRediT).

- Participation in methodological developments and illustrative examples.

5 Scientific papers in the thesis

The scientific papers that make up the thesis are presented there in the following order:

1. Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P. (2016) *Simplicial principal component analysis for density functions in Bayes spaces.* Computational Statistics and Data Analysis 94, 330–350.

2. Talská, R., Hron, K., Matys Grygar, T. (2021) *Compositional scalar-on-function regression with application to sediment particle size distributions.* Mathematical Geosciences 53, 1667–1695.

3. Pavlů, I., Menafoglio, A., Bongiorno, E.G., Hron, K. (2023) *Classification of probability density functions in the framework of Bayes spaces: methods and applications.* Statistics and Operations Research Transactions 47, 295–322.

4. Talská, R., Menafoglio, A., Hron, K., Egozcue, J.J., Palarea-Albaladejo, J. (2020) *Weighting the domain of probability densities in functional data analysis.* Stat 9, e283.

5. Hron, K., Machalová, J., Menafoglio, A. (2023) *Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation.* Statistical Papers 64, 1629–1667.

6. Genest, C., Hron, K., Nešlehová, J.G. (2023) *Orthogonal decomposition of multivariate densities in Bayes spaces and relation with their copula-based representation.* Journal of Multivariate Analysis 198, 105228.

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 44(2):139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

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

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research*, 65(12):4185–4193.

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

Eaton, M. (1983). *Multivariate Statistics. A Vector Space Approach*. John Wiley & Sons.

Eckardt, M., Mateu, J., and Greven, S. (2024). Generalized functional additive mixed models with (functional) compositional covariates for areal covid-19 incidence curves. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73:880–901.

Egozcue, J., Pawlovsky, V., Templ, M., and Hron, K. (2015). Independence in contingency tables using simplicial geometry. *Communications in Statistics*, 44(18):3978–3996.

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

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

Egozcue, J. J. and Pawlowsky-Glahn, V. (2018). Evidence functions: A compositional approach to information. *SORT-Statistics and Operations Research Transactions*, 42(2):1–24.

Fačevicová, K., Filzmoser, P., and Hron, K. (2022). Compositional cubes: a new concept for multi-factorial compositions. *Statistical Papers*, 64:955–985.

Fačevicová, K., Hron, K., Todorov, V., and Templ, M. (2018). General approach to coordinate representation of compositional tables. *Scandinavian Journal of Statistics*, 45:879–899.

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied compositional data analysis*. Springer, Cham.

Genest, C., Hron, K., and Nešlehová, J. (2023). Orthogonal decomposition of multivariate densities in bayes spaces and relation with their copula-based representation. *Journal of Multivariate Analysis*, 198:105228.

Greenacre, M. (2018). *Compositional data analysis in practice*. CRC Press, Boca Raton.

Hron, K., Machalová, J., and Menafoglio, A. (2023). Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation. *Statistical Papers*, 64:1629–1667.

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

Kokoszka, P. and Reimherr, M. (2015). *Introduction to functional data analysis*. CRC Press, Boca Raton.

Kutta, T., Jach, A., Haddad, M., Kokoszka, P., and Wang, H. (2025). Detection and localization of changes in a panel of densities. *Journal of Multivariate Analysis*, 205:105374.

Lei, X., Chen, Z., and Li, H. (2023). Functional outlier detection for density-valued data with application to robustify distribution-to-distribution regression. *Technometrics*, 65(3):351–362.

Ma, Y., Zhou, X., and Wu, W. (2024). A stochastic process representation for time warping functions. *Computational Statistics and Data Analysis*, 194:107941.

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

Machalová, J., Talská, R., Hron, K., and Gába, A. (2021). Compositional splines for representation of density functions. *Computational Statistics*, 36:1031–1064.

Matys Grygar, T., Radojičić, U., Pavlů, I., Greven, S., Nešlehová, J., Tůmová, Š., and Hron, K. (2024). Exploratory functional data analysis of multivariate densities for the identification of agricultural soil contamination by risk elements. *Journal of Geochemical Exploration*, 259:107416.

Murph, A., Strait, J., Moran, K., Hyman, J., and Stauffer, P. (2024). Visualisation and outlier detection for probability density function ensembles. *Stat*, 13:e662.

Pavlů, I., Machalová, J., Tolosana-Delgado, R., Hron, K., Bachmann, K., and van den Boogaart, K. (2024). Principal component analysis for distributions observed by samples in Bayes spaces. *Mathematical Geosciences*, 56:1641–1669.

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Wiley, Chichester.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 15(5):384–398.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, LX:489–502.

Petersen, A., Zhang, C., and Kokoszka, P. (2022). Modeling probability density functions as data objects. *Econometrics and Statistics*, 21(C):159–178.

Podlešáková, E., Němeček, J., and Hálová, G. (1996). Proposal of soil contamination limits for persistent organic xenobiotic substances in the czech republic. *Rostlinna Vyroba*, 42(2):49–54.

Poláková, Š., Hutařová, K., Reininger, D., and Kubík, L. (2011). Registr kontaminovaných ploch 2 M HNO3 (1990 – 2009). Technical report, "Ústřední kontrolní a zkušební ústav zemědělský v Brně", "Brno, Czech Republic".

Qiu, J., Dai, X., and Zhu, Z. (2024). Nonparametric estimation of repeated densities with heterogeneous sample sizes. *Journal of the American Statistical Association*, 119(545):176–188.

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

Scealy, J. L. and Welsh, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 73(3):351–375.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231.

Talská, R., Hron, K., and Matys Grygar, T. (2021). Compositional scalar-on-function regression with application to sediment particle size distributions. *Mathematical Geosciences*, 53:1667–1695.

Talská, R., Menafoglio, A., Hron, K., Egozcue, J. J., and Palarea-Albaladejo, J. (2020). Weighting the domain of probability densities in functional data analysis. *Stat*, 9(1):e283.

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

Tsagris, M. and Stewart, C. (2018). A Dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, 39:398–412.

van den Boogaart, K. and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Springer, Heidelberg.

van den Boogaart, K.-G., Egozcue, J. J., and Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *Statistics and Operations Research Transactions*, 34(2):201–222.

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

Zbíral, J., Honsa, I., Malý, S., and Čižmár, D. (2004). *Soil Analysis III*. Central Institute for Supervising and Testing in Agriculture, Brno, Czech Republic.

Resumé

Bayes spaces provide a general framework for representing relative (distributional) data, covering compositional data and their multifactorial generalisation, as well as univariate and multivariate probability density functions (PDFs). The key property is the scale invariance of these data objects (as well as their associated measures), meaning that any positive constant multiple of them carries essentially the same information. The Hilbert space structure of Bayes spaces allows to consider a general reference measure for domain weighting, as well as to define an isometric isomorphism with the standard geometric framework of multivariate (functional) data, the so-called centred logratio transformation, to enable the use of popular methods of multivariate statistics and functional data analysis, respectively.

The thesis focuses on the analysis of samples of probability density functions (density data) using the Bayes space methodology. Its aim is to summarise the main achievements into a concise methodology for density data analysis, where the Candidate has contributed significantly. Specifically, the efforts can be characterised as twofold: (1) building a concise methodology for density data processing by adapting popular methods of functional data analysis, (2) contributing to the theoretical development of Bayes spaces themselves. There is a record of scientific publications for both. In (1), dimension reduction of a sample of PDFs using simplicial functional principal component analysis, scalar-on-function regression with density as a functional covariate, and classification of PDFs using adapted popular methods of functional data analysis were developed. In (2), weighted density data analysis was proposed by an appropriate choice of the reference measure; understanding the structure of the Bayes space then allowed the development of bivariate and, in the sequel, multivariate extensions of Bayes spaces.

The scientific papers included in the thesis are accompanied by other references that demonstrate the Candidate's extensive track record in analysing relative data. And all of them aim to clearly demonstrate the strong potential of the Bayes space methodology to address challenges in the statistical processing of distributional data.