WORKING PAPERS

N° 1500

October 2024

"The Stick-Breaking and Ordering Representation of Compositional Data: Copulas and Regression models"

Olivier Faugeras

Toulouse School of Economics

The Stick-Breaking and Ordering Representations of Compositional Data: Copulas and Regression models

Olivier P. Faugeras[a,*]

[a] *Toulouse School of Economics, Université Toulouse 1 Capitole, 1 Esplanade de l'université, 31080 Toulouse Cedex 06, France.*

Abstract

Compositional Data (CoDa) is usually viewed as data on the simplex and is studied via a log-ratio analysis, following the classical work of Aitchison [2]. We propose to bring to the fore an alternative view of CoDa as a stick breaking process, an approach which originates from Bayesian nonparametrics. The first stick-breaking approach gives rise to a view of CoDa as ordered statistics, from which we can derive "stick-ordered" distributions. The second approach is based on a rescaled stick-breaking transformation, and give rises to a geometric view of CoDa as a free unit cube. The latter allows to introduce copula and regression models, which are useful for studying the internal or external dependence of CoDa. These stick-breaking representations allow to effectively and simply deal with CoDa with zeroes. We establish connections with other topics of probability and statistics like i) spacings and order statistics, ii) Bayesian nonparametrics and Dirichlet distributions, iii) neutrality, iv) hazard rates and the product integral, v) mixability.

Keywords: Compositional data analysis, Stick-breaking representation, Copula, Regression, Distribution, Order statistic.
2020 MSC: Primary 62H05, Secondary 62G08

1. Introduction and outline

1.1. A primer on CoDa

Compositional Data (CoDa) analysis deals with statistical analysis of d-variate data which are quantitative descriptions of the parts of some whole, conveying only relative information. Composition of soil in geology, elements in a mixture in chemistry, sources of calories in nutrition, vote shares in an election, microbiome data in biology, or a portfolio of financial assets are examples of CoDa.

There are several competing ways to describe CoDa. The traditional approach ([2], [7]) considers that one of the key characteristics of CoDa is that the sum of the proportions must always be equal to a constant (w.l.o.g. 1). This means that the different components of a CoDa point are often expressed as percentages or fractions, rather than absolute values. Hence, Aitchison's approach normalizes a raw composition vector by its sum (an operation called closure in the CoDa literature): let $\mathbf{y} = (y_1, \dots, y_d) \in \mathbb{R}^d_{\geq 0}$ be a vector of non-negative absolute values of a composition, its closure is denoted as

$$C(\mathbf{y}) := \frac{\mathbf{y}}{\|\mathbf{y}\|_1} = \frac{\mathbf{y}}{\sum_{i=1}^d y_i}.$$

This leads to the consideration of *normalised*[1] (i.e. after rescaling to unit sum) CoDa element as a vector $\mathbf{p} = (p_1, \dots, p_d)$, taking its values in the $d-1$ dimensional unit simplex[2]

$$\Delta^{d-1} := \{\mathbf{p} \in \mathbb{R}^d : p_i \geq 0, \sum_{i=1}^d p_i = 1\}. \tag{1}$$

[*] Email address: `olivier.faugeras@tse-fr.eu`

[1] For an intrinsic approach to CoDa analysis based on projective geometry, see [15].

[2] Note that we allow CoDa points with some null components.

Being in a compact space, CoDa elements of Δ^{d-1} can not all be mapped via homeomorphisms to $\mathbb{R}^{d-1}$ and thus do not enjoy a *global* vector space structure. This, and the spurious correlation effect ([37]) induced by the closure operation, prevents the direct application of classical multivariate statistical analysis techniques (e.g. [3]) to the simplex.

Aitchison's ([1], [2]) seminal approach is to study CoDa through a variety of log-ratio transforms: for example, the Additive Log-Ratio transformation (alr) of [2]

$$y_i := \log(p_i/p_d), \quad i = 1, \ldots, d-1$$

and its variants clr and ilr ([23], [36]) turns a CoDa point $\mathbf{p}$ into a *vector* element of $\mathbb{R}^{d-1}$. This gives rise to a special geometry, called Aitchison geometry, which turns the *positive* simplex $\mathring{\Delta}^{d-1}$ into an Euclidean vector space. For recent accounts on the latter, see e.g., the survey article [24], and the books [2], [23], [36], [19], [7].

However, due to the log, log-ratios are undefined if $\mathbf{p}$ has some zero components[3]. Thus, the above-mentioned statistical literature based on log-ratio analysis usually enforces a strict non-negativity assumption of the CoDa components, which limits its scope of application. (Special treatments of the zero components are required: these range from ad-hoc methods like amalgamation of the finer parts, replacement of the zeroes with small values, treatment of the zero observations as outliers to more involved approaches based on the stratification of the simplex according to the zeroes patterns.) This motivates the search of possible alternative representations of CoDa on the *full* simplex Δ^{d-1}.

1.2. Aims and scope

The purpose of this paper is to bring to the attention of the CoDa community a possible simple alternative approach for the statistical analysis of CoDa. It is inspired and finds its origin in Bayesian non-parametrics and consider CoDa as a stick-breaking process. This give rises to two interrelated geometric views on CoDa points. The first is that of an ordered set on the unit interval. This ordered view allows to define distributions on CoDa points via order statistics and spacings of a latent vector on the unit interval.

The second view is based on a transformation which gives the relative positions of the breaks, yielding a parametrization of the CoDa point as an unconstrained unit cube. These relative positions have an interpretation as conditional probabilities and are related to the concept of neutrality, a natural intra-independence notion for compositions. It can also be apprehended in terms of hazard rate, via the product integral. This second view also allows to define distributions on the simplex, in particular copula distributions, for the study of the intra-dependence of CoDa. In addition, it is useful to study internal (resp. external) regression models, i.e. when one wants to explain/predict a (set of) components by other components acting as predictors (resp. by external covariates).

As these stick-breaking approaches can be described from multiple viewpoints and have multiple origins, our second aim is thus to survey these and show the connections that exist between several topics of probability and statistics, like spacings and order statistics, Bayesian nonparametrics and Dirichlet distributions, neutrality and subcompositions, hazard rates and the product integral, and mixability.

1.3. Outline

The outline of the paper is as follows: in Section 2, we introduce the first stick breaking transformation and define stick-breaking distributions for CoDa, based on spacings and order statistics. Several examples are given and numerically illustrated. Section 3 elaborates on the first construction by considering a rescaled version of the stick-breaking process. It turns the constrained CoDa point of the simplex into a free vector of the unit cube $[0, 1]^{d-1}$, which can, for positive CoDa, even be transformed to a free Euclidean vector of $\mathbb{R}^{d-1}$. These approaches yield a triple representation of CoDa. We show the connections with neutrality, Dirichlet distributions, hazard rate and iterative partitioning, and briefly discuss some possible variants.

Section 4 and 5 deal with statistical applications of such rescaled stick-breaking transformation for the study of the intra-dependence of CoDa. Section 4 introduces CoDapulas as the analogue of copulas for CoDa, opening the gates of the vast copula literature, tools and methodologies for CoDa. Several examples illustrates how copula models can easily be constructed for CoDa. In particular, completely monotone copulas give interesting complete dependence

[3] Ratios are also undefined if both numerator and denominator are zero. See the forthcoming Proposition 4 for a treatment of this case.

patterns for CoDa. Section 5 aims at studying intra-dependence of CoDa from the regression viewpoint. A basic example of a parametric regression model on real data set is given. Several extensions and alternatives are discussed. Eventually, we conclude in Section 6, with additional remarks about the choice of ordering of the components, and mixability.

2. The first stick-breaking view on CoDa: ordered points on the unit interval

2.1. The ordered stick-breaking view

The first stick-breaking approach is based on the representation of CoDa as a normalised point $\mathbf{p}$ in the simplex (1). Instead of considering the (p_i), $1 \leq i \leq d$, as primary parameters for $\mathbf{p}$, one can consider the accumulated sums $\mathbf{s} = (s_0, s_1, \ldots, s_d)$, defined as

$$\begin{aligned} s_0 &= 0, \\ s_i &= p_i + s_{i-1} = \sum_{j=1}^{i} p_j, \quad i = 1, \ldots, d-1, \\ s_d &= \sum_{i=1}^{d} p_i = 1, \end{aligned} \tag{2}$$

as an alternative system of coordinates of (1). Dropping the fixed values $s_0 = 0$ and $s_d = 1$, this leads to a representation of (1) as

$$\Sigma^{d-1} := \{(s_1, \ldots, s_{d-1}) \in \mathbb{R}^{d-1} : 0 \leq s_1 \leq \ldots \leq s_{d-1} \leq 1\}. \tag{3}$$

Equation (3) can be interpreted as iteratively breaking the unit stick $[0, 1]$: first, one picks some $s_1 \in [0, 1]$, then, for the second step, one picks some s_2 in the remaining interval $[s_1, 1]$, and so on for s_i to be picked in the interval $[s_{i-1}, 1]$, $i = 1, \ldots, d-1$. The process stops after $d-1$ steps. See Figure 1.

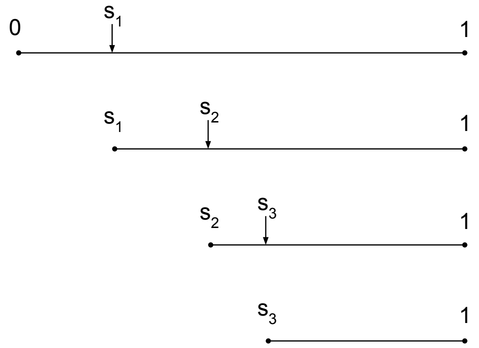

Fig. 1: Stick-breaking of the unit interval: each remaining interval $[s_i, 1]$ is broken at $s_{i+1} \in [s_i, 1]$.

Conversely, $d-1$ ordered values $(s_1, \ldots, s_{d-1}) \in \Sigma^{d-1}$ determines uniquely a CoDa point $\mathbf{p}$ in the simplex Δ^{d-1} by

$$\begin{aligned} p_1 &= s_1, \\ p_i &= s_i - s_{i-1}, \quad i = 2, \ldots, d-1, \\ p_d &= 1 - s_{d-1}. \end{aligned} \tag{4}$$

For the record, let us formalize these elementary considerations in the following:

Proposition 1. *The summing transformation $\mathcal{S} : \mathbf{p} \mapsto \mathbf{s}$ defined by (2) is a bijection from Δ^{d-1} to Σ^{d-1}, with inverse transformation given by (4).*

Remark 1.

1. A normalized CoDa point $\mathbf{p} \in \Delta^{d-1}$ can be identified with a discrete probability measure $\mu_{\mathbf{p}}$ on $\mathbb{R}$,

$$\mathbf{p} \hookrightarrow \mu_{\mathbf{p}}(.) := \sum_{i=1}^{d} p_i \delta_{x_i}(.),$$

where $x_1 < \ldots < x_d \in \mathbb{R}$ denotes (arbitrarily located) distinct components and δ_x stands for the Dirac mass at x. The Coda $\mathbf{p}$ "forgets" about the locations $x_1, \ldots, x_d$ of $\mu_{\mathbf{p}}$ of the components, to only retains their probabilities $p_1, \ldots, p_d$. As such, parametrization (2) of the simplex by the (s_i) interprets as characterizing the discrete distribution of a r.v. $X \sim \mu_{\mathbf{p}}$ by its c.d.f. $F(x) = P(X \leq x) = \sum_{y \leq x} P(X = x)$, while the parametrization by the (p_i) interprets as characterizing the distribution of X by its probability mass function $P(X = x)$.
2. (3) only uses the order structure of the interval $[0, 1]$, and not the addition operation. This suggests that one can generalise the notion of simplex to arbitrary ordered space, endowed with a top (= 1) and bottom (= 0) element.
3. Instead of breaking the stick from the left to the right, i.e. putting $s_i \in [s_{i-1}, 1]$, for increasing $i = 1, \ldots, d-1$, one can also consider a stick-breaking process from the right to the left, i.e. putting $s_i \in [0, s_{i+1}]$ for decreasing $i = d-1, \ldots, 1$. This corresponds to characterizing $\mu_{\mathbf{p}}$ by its survival function instead of its cumulative distribution function.

2.2. *CoDa distributions via order statistics*

2.2.1. *Stick-Ordered distributions*

This geometric view of CoDa as a set of ordered points on the unit interval suggests a natural connection with order statistics on the unit interval. This gives an easy way to build distributions on the simplex by taking as s_i the order statistics of some u_i distributed on the unit interval. More precisely, one can define a "Stick-Ordered" (SO) distribution on the simplex as follows:

Definition 2 (Stick-Ordered distribution for CoDa)**.** Let $u_i \sim F_i$, $i = 1, \ldots d-1$ be independent r.v. with $(F_1, \ldots, F_{d-1})$ a set of univariate c.d.f.s on the unit interval. Set

$$u_{(1)} \leq \ldots \leq u_{(d-1)}$$

the corresponding order statistics. Eventually, define

$$s_0 = 0, \quad s_i = u_{(i)}, \; i = 1, \ldots d-1, \quad s_d = 1.$$

Then, the CoDa point $\mathbf{p} \in \Delta^{d-1}$ corresponding to $\mathbf{s}$ in (3) is said to be $(F_1, \ldots, F_{d-1})$-Stick-Ordered distributed, which is denoted by

$$\mathbf{p} \sim \mathtt{SO}(F_1, \ldots, F_{d-1}).$$

In case $F_i = F$, $\mathbf{p}$ is said to be F-Stick-Ordered distributed, which is denoted by $\mathbf{p} \sim \mathtt{SO}(F)$.

In other words, the Stick-Ordered distribution of $\mathbf{p}$ is the distribution of the spacings corresponding to the order statistics $\mathbf{s}$. The latter has to be computed as the distribution of (possibly non-identically distributed) order statistics.

$$\mathbf{p} \sim \mathtt{SO}(F_1, \ldots, F_{d-1}) \Longleftrightarrow \begin{cases} u_i \sim F_i, \quad (u_1, \ldots, u_{d-1}) \text{ independent,} \\ s_0 = 0, s_i = u_{(i)}, s_d = 1, \quad i = 1, \ldots, d-1, \\ p_i = s_i - s_{i-1}, \quad i = 1, \ldots, d \end{cases}$$

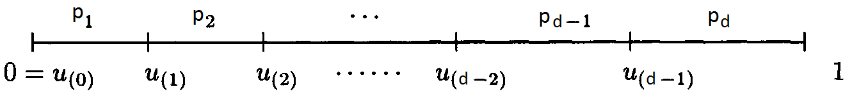

Fig. 2: Stick-Ordered distribution $\mathtt{SO}$, obtained from the order statistics $u_{(i)}$: each p_i corresponds to a spacing.

2.2.2. Examples

Example 1. *In particular, for $F = U_{[0,1]}$ the uniform distribution, (i.e. $F(x) = x,\ 0 < x < 1$), $\mathbf{p} \sim \mathcal{SO}(U_{[0,1]})$ gives a uniform distribution on the simplex, as shown in [38] equation (2.1). (See also [50] equation (6)). More precisely, $(p_1, \ldots, p_{d-1})$ has a density w.r.t. the $d-1$ dimensional Lebesgue measure given by*

$$f_{(p_1,\ldots,p_{d-1})}(x_1, \ldots, x_{d-1}) = \begin{cases} (d-1)! & \textit{if } x_i \geq 0, \textit{ and } \sum_{i=1}^{d-1} x_i \leq 1 \\ 0 & \textit{otherwise} \end{cases}.$$

On the other hand, $\mathbf{p}$ has a singular distribution since $p_d = 1 - \sum_{i=1}^{d-1} p_i$, but its restriction to the hyperplane $\sum_{i=1}^{d} p_i = 1$ admits a density w.r.t. the $d-1$ dimensional Lebesgue measure given by

$$f_{\mathbf{p}}(\mathbf{x}) = \begin{cases} (d-1)! & \textit{if } x_i \geq 0, \textit{ and } \sum_{i=1}^{d} x_i = 1 \\ 0 & \textit{otherwise} \end{cases},$$

which is symmetric in $(x_1, \ldots, x_d)$ (i.e. the (p_i) are exchangeable). One recognizes the Dirichlet $\mathcal{Dir}(1, \ldots, 1; 1)$ distribution, see e.g. [39] or [51].

These stick-ordered distributions are useful for modeling purposes. They allow to construct CoDa models from classical distributions on $[0, 1]$. One can consider more examples with other distributions on the unit interval, like the Beta, the Kumaraswamy, (which is similar to the Beta distribution but leads to tractable formulas for the distribution of order statistics, see [30]), or those of [31]. (More generally, any distribution on $\mathbb{R}$ can be mapped to a distribution with support included on the unit interval by applying a c.d.f to it). In some cases, analytical formulas can be obtained for the distribution of $\mathbf{p}$, using known results on spacings ([38]).

Example 2 (Kumaraswamy)**.** *The Kumaraswamy distribution ([30]) $U \sim$ Kumaraswamy(α,β) has density*

$$f(u) = \alpha\beta u^{\alpha-1}(1 - u^{\alpha})^{\beta-1}, \quad 0 < u < 1, \tag{5}$$

and cdf

$$F(u) = 1 - (1 - x^{\alpha})^{\beta}, \quad 1 < u < 1. \tag{6}$$

Moreover, for i.i.d. $u_i \sim F$, $i = 1, \ldots, d-1$, with density f, the marginal distribution of the spacings writes (see e.g. [38] p. 399)

$$f_{p_i}(x) = \frac{(d-1)!}{(i-2)!(d-1-i)!} \int (F(t))^{i-2}(1 - F(x+t))^{d-1-i} f(t) f(x+t) dt, \tag{7}$$

for $2 \leq i \leq d-1$, and

$$f_{p_1}(x) = f_{u_{(1)}}(x) = (d-1) f(x)(1 - F(x))^{d-2}. \tag{8}$$

Applying formulas (7) and (8) to (5) and (6) leads to computable formulas. For example,

$$f_{p_1}(x) = (d-1)\alpha\beta x^{\alpha-1}(1 - x^{\alpha})^{\beta(d-1)-1}, \quad 0 < x < 1,$$

The Markov property of the order statistics can then be used to derive the joint distribution of $\mathbf{p}$.

2.2.3. Generalised Stick-Ordered Distributions

One can also generalise the former definition (2) by taking dependent r.v. u_i instead of independent ones.

Definition 3 (Generalized stick-ordered distribution for CoDa)**.** For $\mathbf{u} = (u_1, \ldots, u_{d-1}) \in [0, 1]^{d-1}$ with joint distribution function $F_{\mathbf{u}}$, the CoDa point $\mathbf{p} \in \Delta^{d-1}$ corresponding to $\mathbf{s}$ in (3) is said to be generalized-stick-ordered distributed with generator $F_{\mathbf{u}}$, which is denoted by $\mathbf{p} \sim \texttt{GSO}(F_{\mathbf{u}})$, viz.

$$\mathbf{p} \sim \texttt{GSO}(F_{\mathbf{u}}) \iff \begin{cases} (u_1, \ldots, u_{d-1}) \sim F_{\mathbf{u}} \\ s_0 = 0, s_i = u_{(i)}, s_d = 1, \quad i = 1, \ldots, d-1, \\ p_i = s_i - s_{i-1}, \quad i = 1, \ldots, d \end{cases}$$

For non-identically distributed or dependent variables, one can compute them e.g. using results of [10] Chap. 5. [29], [43] (See also [5]). However, this often leads to intractable formulas. Nonetheless, it is easy to simulate samples from such distributions.

Example 3 (GSO with Gaussian copula generator)**.** *Consider, for $d = 3$, $\mathbf{p} \sim \mathtt{GSO}(F_{\mathbf{u}})$ with $F_{\mathbf{U}}$ a bivariate Gaussian copula (hence with uniform marginals), with correlation ρ. Figure 3 shows ternary diagrams of scatterplots of samples of* 1000 *realisations of $\mathbf{p} = (p_1, p_2, p_3)$, with varying level of the dependence coefficient ρ. The value of ρ determines the behavior of the distribution of the CoDa element* $\mathbf{p}$ *and generates interesting patterns of dependence between the components. For $\rho = 0$, one generates a uniform distribution on the simplex (upper right panel). When ρ becomes negative (upper middle panel) and close to -1 (upper left panel), one obtains empirically a CoDa point s.t. $p_2 \approx 1 - 2p_1$ and $p_1 \approx p_3$. For ρ close to one ($\rho = 0.99$, lower right panel), the p_2 component is nearly zero and the CoDa point is nearly on the line $p_1 = p_3$.*

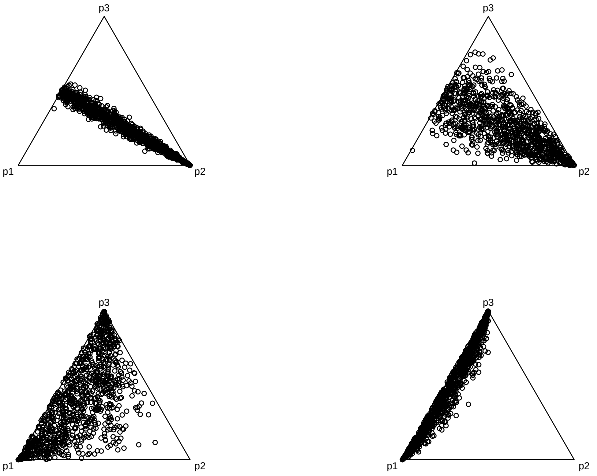

Fig. 3: Ternary plots of Generalized-Stick-Ordered distribution for $d = 3$, with Gaussian copula generator, with varying correlation coefficient ρ. From left to right and up to down: $\rho = -0.99, \rho = -0.8, \rho = 0, \rho = 0.8, \rho = 0.99$.

3. The rescaled stick-breaking view: unit cube geometry of CoDa points

3.1. Unit cube geometry for CoDa points by rescaling

The second approach we promote is based on a rescaled version of the iterative stick-breaking process of Figure 1: first, one picks some $s_1 \in [0, 1]$, as previously. Then, one has to pick s_2 in the remaining interval $[s_1, 1]$: in terms of spacings/lengths, the length of $s_2 - s_1 = p_2$ of the second stick $[s_1, s_2]$ has to be chosen relatively to the length $1 - s_2$ of the remaining stick $[s_2, 1]$, see Figure 2. Similarly, the relative length $s_i - s_{i-1} = p_i$ of the interval corresponding to the ith pick s_i has to be chosen relatively to the length $1 - s_{i-1}$ of the remaining stick $[s_{i-1}, 1]$. Following the footsteps of [27], [9] among others, it is therefore natural to introduce the transformation,

$$\begin{aligned} z_1 &= s_1 = p_1, \\ z_i &= \frac{p_i}{1 - s_{i-1}}, \quad i = 1, \ldots, d - 1, \\ z_d &= 1, \end{aligned} \tag{9}$$

with the convention that $0/0 := 0$.

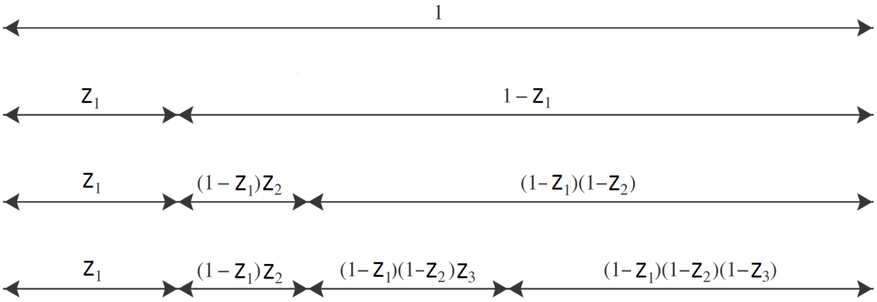

Fig. 4: Rescaled stick-breaking: each remaining stick is broken, relatively to its length, by $z_i \in [0, 1]$.

By construction, since $1 - s_{i-1} = p_i + \ldots + p_d \geq p_i$, the z_i in (9) are in the unit cube, $0 \leq z_i \leq 1$ for $i = 1, \ldots, d-1$, with degenerate $z_d = 1$. Thus the transformation (9) turns the "akward" simplex Δ^{d-1} into the unit cube $[0, 1]^{d-1}$ (we drop z_d as it is always equal to 1). This leads to a free unit cube view of CoDa points as an unconstrained element of $[0, 1]^{d-1}$. More precisely, the transformation (9) realizes "almost" a bijection of the simplex, as formalized in the next proposition:

Proposition 4. *Let the rescaled stick-breaking transformation*

$$\mathcal{R} : \Delta^{d-1} \to [0, 1]^{d-1}$$
$$\mathbf{p} \mapsto \mathbf{z}$$

be defined by (9). Then,

i) $\mathcal{R}$ is a bijection from the interior of the simplex $\mathring{\Delta}^{d-1}$ to the open cube $(0, 1)^{d-1}$, with inverse transformation $\mathcal{R}^{-1}$ given by $\mathbf{p} = \mathcal{R}^{-1}(\mathbf{z})$ with

$$\begin{aligned} p_1 &= z_1, \\ p_i &= z_i \prod_{j=1}^{i-1} \left(1 - z_j\right), \quad i = 2, \ldots, d-1, \\ p_d &= \prod_{i=1}^{d-1} (1 - z_i). \end{aligned} \tag{10}$$

ii) $\mathcal{R}^{-1} : [0, 1]^{d-1} \to \Delta^{d-1}$ is a retraction (left-inverse) of $\mathcal{R}$ on the full simplex Δ^{d-1}, i.e. $\mathcal{R}^{-1} \circ \mathcal{R}(\mathbf{p}) = \mathbf{p}$, for all $\mathbf{p} \in \Delta^{d-1}$.

Proof. $\mathcal{R}$ is ill-defined when the denominator in (9) is zero, that is to say when there exists some $1 \leq i \leq d-2$ s.t. $s_i = 1$. Notice that if this so, then $s_j = 1$ and $p_j = 0$, for $j > i$. This entails that $z_i = p_i/p_i = 1$ and $z_j = 0$ for $j > i$, with our convention that $0/0 := 0$. Conversely, if $z_i = 1$ for some $1 \leq i \leq d-2$, then $p_{i+1} + \ldots + p_d = 0$ which entails $p_j = z_j = 0$ for $j > i$.

So, let i_0 the smallest i such that $s_i = 1$, $1 \leq i \leq d$. If $\mathbf{p} \in \mathring{\Delta}^{d-1}$ or $\mathbf{z} \in (0, 1)^{d-1}$, then, in view of the preceding, $i_0 > d-2$, $\mathcal{R}$ is well-defined and simple algebraic manipulations show that $\mathcal{R}$ and $\mathcal{R}^{-1}$ are inverse of each other. This settles case i). If $1 \leq i_0 \leq d-2$, $\mathbf{p} = (p_1, \ldots, p_{i_0}, 0, \ldots, 0)$, $\mathbf{z} = \mathcal{R}(\mathbf{p}) = (z_1, \ldots, z_{i_0-1}, 1, 0, \ldots, 0)$ and $\mathcal{R}^{-1}(\mathbf{z}) = \mathbf{p}$. Thus, in the general case ii), for $\mathbf{p} \in \Delta^{d-1}$, $\mathcal{R}^{-1} \circ \mathcal{R}(\mathbf{p}) = \mathbf{p}$.

□

In other words, $\mathbf{z}$ can only have one of its coordinates equal to 1 (with remaining coordinates equal to zero, if so happens). This implies that the faces of the cube $[0, 1]^{d-1}$ with more than one 1 in their coordinates do not correspond to a CoDa point: some parts of the boundary of the cube $[0, 1]^{d-1}$ can not be mapped back to the simplex. Nonetheless, this is not a serious restriction. Proposition 4 ii) means that the whole simplex can be injected into the unit cube and mapped back to the simplex: all CoDa points (including CoDa with zeroes) can be studied in their $\mathbf{z}$ coordinates, as points of (almost all) the unit cube.

Remark 2. Historically, the transformation (9) was introduced by [27] for $d = \infty$, as a way to distribute gold dust to a countably infinite sequence of beggars, where each beggar receives in turn a fraction z_i of the remaining gold. More generally, an infinite sequence $(p_1, p_2, \ldots)$ defined by (9) from a sequence of independent $(z_i) \in [0, 1]$ is called a residual allocation model (RAM), see [17] for a review. The case when $z_i \sim \beta(1, \theta)$, $\theta > 0$ is called a GEM distribution and was notably studied in population genetics by [26], [12], [32], see e.g. [14]. RAM also appears in Bayesian statistics in connection with the Dirichlet distribution, as the weights of random measures $\mathcal{P}(.) = \sum_{k=1}^{\infty} p_k \delta_{\xi_k}(.)$, where (ξ_k) are i.i.d. and independent of (p_k), see [51] p. 178, [18], [28]. See also Section 3.3 below.

3.2. Neutrality and complete neutrality

This interpretation of CoDa points as a relative/proportional iterative stick-breaking process leads to the concept of neutrality, introduced by [9], which is relevant for the analysis of CoDa. In short, it is a sort of intra-independence concept for a random composition p.

More precisely, neutrality is motivated by the following: if one wants to check whether the first proportion p_1 has an influence on the remaining subcomposition $(p_2, \ldots, p_d)$, the latter has to be rescaled by the remaining mass $1 - p_1$, in order to be a proper normalized CoDa point. One thus has to check for the stochastic influence of p_1 on

$$\left(\frac{p_2}{1-p_1}, \quad \ldots, \quad \frac{p_d}{1-p_1}\right), \tag{11}$$

and if p_1 is independent of the latter rescaled subcomposition, one can eliminate p_1 from the analysis of $\mathbf{p}$. Therefore, [9] defines neutrality as follows: p_1 is said to be neutral if p_1 is independent of (11): p_1 does not influence the manner in which the remaining proportions $(p_2, \ldots, p_d)$ relatively divide the remainder of the unit interval.

A generalisation of neutrality to a vector $\mathbf{p}_k := (p_1, \ldots, p_k)$, $k < d$ is: $(p_1, \ldots, p_k)$ is a neutral vector if it is independent of

$$\left(\frac{p_{k+1}}{1-s_k}, \ldots, \frac{p_d}{1-s_k}\right).$$

Thus, if $\mathbf{p}_j$ is neutral for $j = 1, \ldots k$, then $\mathbf{z}_k := (z_1, \ldots, z_k)$ is mutually independent (Theorem 1 in [9]). A further generalization of neutrality is complete neutrality: if the $\mathbf{z} = (z_1, \ldots, z_{d-1})$ of (9) are mutually independent, then the corresponding $\mathbf{p}$ is said to be completely neutral, or equivalently (Theorem 2 in [9]) if $\mathbf{p}_j$ is neutral for all $1 \leq j \leq d-1$.

These concepts of neutrality are helpful for constructing completely neutral distributions on the simplex: start with mutually independent z_i's each having a specified distribution on $[0, 1]$, and invert (9) to obtain a completely neutral distribution on the simplex. In particular, [9] construct a generalisation of the Dirichlet distribution from independent $z_i \sim \beta(a_i, b_i)$. [35] Theorem 2.2 use the transformations (9) and (10) to obtain stochastic representations of the Dirichlet distribution $D(\mathbf{a})$ from independent $z_i \sim \beta(a_i, \sum_{k=i+1}^{d} a_k)$, $i = 1, \ldots, d-1$.

3.3. Conditional probability interpretation of the rescaled stick-breaking approach and connection with Bayesian priors

The rescaled weights z_i interpret as conditional probabilities. The stick-breaking construction appears in the construction of the (finite-dimensional) Dirichlet distribution of [18], see e.g. [21] p. 30. The latter is used for constructing a prior on a discrete distribution in Bayesian statistics. It is defined as follows: in order to randomly distribute a total mass 1, identified with the unit interval, to the first d integers $1, 2, \ldots, d$, the stick is first randomly broken by a r.v. $0 \leq Z_1 \leq 1$, and mass Z_1 is assigned to 1. The remaining mass is $1 - Z_1$ and the stick $[Z_1, 1]$ is broken into two new pieces of *relative* length Z_2 and $1 - Z_2$, for some $0 \leq Z_2 \leq 1$. Mass $(1 - Z_1)Z_2$ is assigned to the point 2, and the remaining stick has remaining mass (or length) $(1 - Z_1)(1 - Z_2)$. Iterating, one has defined a random distribution (i.e. a Markov kernel), with values $j = 1, \ldots, d$ and (random) probabilities given by (10).

Each p_i is the probability assigned to i, conditionally on the previous probabilities assigned to the $j < i$. Indeed, if one denotes by ζ the r.v. with values in $1, \ldots, d$ and (random) probabilities given by (10), i.e. s.t.

$$P(\zeta = i) = p_i = Z_i \prod_{j=1}^{i-1} \left(1 - Z_j\right), \quad i = 1, \ldots, d,$$

Then, $Z_i = P(\zeta = i | \zeta \geq i)$.

On the other hand, the complete neutrality property expresses the idea that these Z_i (or equivalently these conditional probabilities) are chosen independent. In particular, if $(Z_1, \dots, Z_d)$ are independent with $Z_i \sim \beta(\alpha_i, \sum_{j>i} \alpha_j)$, then $\mathbf{p}$ is Dirichlet $\mathtt{Dir}(k; \alpha_1, \dots, \alpha_d)$ distributed, see [21] Corollary G.5.

3.4. Hazard rate interpretation with the product integral

Going one step further in the interpretation of the rescaled stick-breaking transform (9) in terms of conditional probabilities of Section 3.3, we now show that it is associated with the expression of density functions in terms of hazard rate, a concept commonly used in reliability and survival analysis. Indeed, for X a positive real-valued random variable with density f and cdf F, recall that the hazard rate function is defined as minus the log derivative of the survival function,

$$h(x) := \frac{f(x)}{1 - F(x)} = -\frac{d \ln (1 - F(x))}{dx} = \lim_{h \downarrow 0} \frac{P(X \leq x + h | X > x)}{h}. \tag{12}$$

It interprets, when X stands for a survival time, as the conditional probability of the failure of the device at age x, given that it did not fail before age x. By direct integration of (12), the survival function (or cdf) expresses in terms of the hazard rate, as

$$1 - F(x) = \exp\left(- \int_0^x h(y) dy \right), \quad x \geq 0. \tag{13}$$

Combining (12) and (13) yields a representation of the density function as

$$f(x) = h(x) \exp\left(- \int_0^x h(y) dy \right). \tag{14}$$

Equations (12), resp. (14), are the continuous analogue of the discrete $\mathbf{p} \to \mathbf{z}$ transform (9), resp. the inverse $\mathbf{z} \to \mathbf{p}$ transform (10). For (9) and (12) this is clear, as, in the discrete case, the density (Radon-Nikodym derivative w.r.t. the counting measure) $f(x) = P(X = x)$ identifies with the CoDa vector of probabilities $\mathbf{p}$, and the cdf $F(x)$ with the accumulated sum vector $\mathbf{s}$, in accordance with Remark 1. The $\mathbf{z}$ coordinates thus interpret as a discrete hazard rate.

For the inverse transform $\mathbf{z} \to \mathbf{p}$, the correspondence is less apparent as the continuous density (14) and discrete inverse transforms (10) appear different at first sight. The analogy is complete when one writes (14) in terms of Volterra's product integral [47]. The product integral is the continuous product analogue of the ordinary (Riemann, Lebesgue, Denjoy, Perron etc...) integral and provide a compact functional way to express the solution to the Cauchy problem for systems of ordinary differential equations (or of integral equations). It finds applications in survival analysis (Kaplan-Meier and Nelson-Aalen estimators), nonlinear systems theory (Péano-Wiener series), Markov processes and semi-martingales, see [11], [45], [22] for more details.

Indeed, given a fixed $x > 0$, let $0 = x_0 < x_1 < \dots < x_m = x$ be a partition of $[0, x]$. Then, by successive conditioning,

$$\begin{aligned} P(X > x) = 1 - F(x) &= P(X > x_0) P(X > x_1 | X > x_0) \dots P(X > x_m | X > x_{m-1}) \\ &= \prod_{i=1}^{m} (1 - P(X \leq x_i | X > x_{i-1})) \end{aligned}$$

By (12), $P(X \leq x_i | X > x_{i-1}) \approx h(x_i) \delta x_i$, for $\delta x_i := x_i - x_{i-1}$ small. Thus, when the partition size goes to zero, viz. $m \to \infty$, $\delta := \max_{1 \leq i \ m} \delta x_i \to 0$, then

$$P(X > x) \approx \prod_{i=1}^{m} (1 - h(x_i) \delta x_i) \to \mathop{\Large\pi}\limits_{0}^{x} (1 - h(y) dy)$$

where the r.h.s is the definition of the product integral, understood as the limit of the finite products of the l.h.s. The final step in the analogy consists in the property that, since $e^{-x} = 1 - x + o(x^2)$, the product integral can also be written as an exponential product integral, see e.g. Theorem 1.7.1 p. 52 in [11]. In turn, it is equal to the exponential of the classical Riemann sum integral:

$$\mathop{\Large\pi}\limits_{0}^{x} (1 - h(y) dy) = \mathop{\Large\pi}\limits_{0}^{x} e^{-h(y) dy} := \lim_{\delta \downarrow 0} \prod_{i=1}^{m} \left(e^{-h(x_i) \delta x_i} \right) = \lim_{\delta \downarrow 0} e^{\sum_{i=1}^{m} -h(x_i) \delta x_i} = e^{-\int_0^x h(y) dy}.$$

In other words, the formula of the inverse **z** transform (10),

$$p_i = z_i \prod_{j=1}^{i-1} \left(1 - z_j\right), \quad i = 2, \ldots, d-1,$$

interprets as the discrete analogue of the density formula (14) in terms of hazard rate, that is

$$f(x) = h(x) \prod_{0}^{x} (1 - h(y)dy). \tag{15}$$

This analogy with the hazard rate and connection with survival and reliability analysis suggest that statistical models (like Cox or frailty models) and techniques of the latter fields could be used for CoDa.

Remark 3. [22] develop a theory of product integration for matrix measures $\boldsymbol{\mu}$ on $]0, \infty]$ which allows to simultaneously handle the continuous and discrete case. In particular, their Definition 4 of the product integral in the univariate commuting case gives

$$\prod_{]0,x]} (1 + d\mu) = \prod_{y \in]0,x]} (1 + \mu(\{y\})) \times \exp\left(\mu^c(]0, x])\right),$$

where μ^c is the absolutely continuous part of the measure μ. This unifies the continuous (15) and discrete (10) case in a single formula.

3.5. A triple representation of CoDa

One thus has a triple representation of the simplex / of normalized CoDa points: the simplex can be represented as Δ^{d-1} with its sum constraint, as the ordered set of points Σ^{d-1} on the unit interval, or a free cube $[0, 1]^{d-1}$ via its rescaled representation in **z** coordinates. Figure 5 shows the different representations as well as the transformations between them (where the arrows between the free unit cube $[0, 1]^{d-1}$ and Σ^{d-1} are obtained by composition of the previous transformations). Note that the ordered and rescaled representations are not canonical, as they depend on the order of enumeration of the components of **p**.

$$\begin{tikzcd}
 & \begin{array}{c}\underline{\text{Unit Cube:}}\\ \underline{\text{Relative representation}}\\ \mathbf{z} \in [0,1]^{d-1}\end{array} \arrow[dl] \arrow[dr] & \\
\begin{array}{c}\underline{\text{Simplex:}}\\ \underline{\text{Unit-sum representation}}\\ \mathbf{p} \in \Delta^{d-1} \subset [0,1]^{d}\end{array} \arrow[ur, bend left, "\text{rescaled stick-breaking}"] \arrow[rr, bend right, "\text{summing:}"'] & & \begin{array}{c}\underline{\text{Interval:}}\\ \underline{\text{Ordered representation}}\\ \mathbf{s} \in \Sigma^{d-1} \subset [0,1]^{d-1}\end{array} \arrow[ul, bend right] \arrow[ll, bend right, "\text{spacings/differencing}"'] \\
\mathbf{y} \in \mathbb{R}^{d}_{\geq 0} \arrow[u, "\text{normalisation } \frac{y}{\|y\|_1}"] & & \mathbf{u} \in [0,1]^{d-1} \arrow[u, "\text{ordering}"']
\end{tikzcd}$$

Fig. 5: A triple representation of CoDa: as a simplex, as an ordered set and as a free unit cube.

In addition, Figure 5 shows how one can obtain the stick-ordered distribution of Definition 2, via ordering of some $u_i \in [0, 1]$ r.v. (lower-right), for $i = 1, \ldots, d-1$. Another way to produce a CoDa point is through closure C, i.e.

normalisation by the sum of nonnegative random variables, for some $\mathbf{y} = (y_1, \dots, y_d) \in \mathbb{R}^d_{\geq 0}$. This is also illustrated in Figure 5, (lower-left).

The figure shed lights on some results and representations of order statistics and constructions of Dirichlet distribution. For example, it is well-known (see [46], [40], [38]) that the order statistics and spacings of i.i.d (u_i) r.v. uniformly distributed on [0, 1], have a representation as a ratio of (sums of) exponential r.v.: Take $y_i \sim Exp(1)$ i.i.d. in Figure 5, then normalisation by the sum gives the p_i which corresponds to spacings, and summing this spacings give the order statistics $s_i = u_{(i)}$,

$$(u_{(i)})_{i=1,\dots,d-1} \overset{d}{=} \left(\frac{\sum_{j \leq i} y_j}{\sum_{j=1}^{d} y_j} \right)_{i=1,\dots,d-1}$$

and

$$(p_i)_{i=1,\dots,d} \overset{d}{=} \left(\frac{y_i}{\sum_{j=1}^{d} y_j} \right)_{i=1,\dots,d}$$

Also, for $y_i \sim \gamma(\alpha_i)$ Gamma distributed, Figure 5 allows to explain and visualize the difference between the Dirichlet distribution on Δ^{d-1} and its ordered version on Σ^{d-1}, see [51] p. 178, 182 and 238.

3.6. From the unit cube to the free Euclidean space $\mathbb{R}^{d-1}$

If $\mathbf{p}$ has no zero components, viz. $0 < p_i < 1$ for all $1 \leq i \leq d$, then $\mathbf{p}$ is sent to the interior $(0, 1)^{d-1}$ of the unit cube $[0, 1]^{d-1}$ by the rescaled stick-breaking transformation (9). In turn, one can then map the open unit-cube representation $\mathbf{z} \in (0, 1)^{d-1}$ of the Coda element $\mathbf{p} \in \Delta^{d-1}$ to a point $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{d-1}) \in \mathbb{R}^{d-1}$ by applying an increasing[4] continuous transformation $q : (0, 1) \to \mathbb{R}$ to each component z_i of $\mathbf{z}$, viz.

$$\xi_i = q(z_i), \quad 1 \leq i \leq d - 1.$$

See Figure 6. Examples of q which come to mind include the probit, logit transform, or any quantile function of a distribution on $\mathbb{R}$ with positive density (hence the notation q).

$$\begin{tikzcd}
\begin{array}{c}\mathbf{p} \in \Delta^{d-1}, \\ \boldsymbol{p} > \mathbf{0}\end{array} \arrow[r, bend left, "\text{rescaled stick-breaking}"] & \mathbf{z} \in (0,1)^{d-1} \arrow[l, bend left] \arrow[r, bend left, "q"] & \boldsymbol{\xi} \in \mathbb{R}^{d-1} \arrow[l, bend left, "q^{-1}"]
\end{tikzcd}$$

Fig. 6: Free Euclidean representation $\boldsymbol{\xi} \in \mathbb{R}^{d-1}$ of positive CoDa $\mathbf{p}$ by (quantile) mapping of its rescaled stick-breaking representation $\mathbf{z}$.

This gives an interesting alternative to the vector space representation provided by Aichison's log-ratio transforms. This variant of the $\mathbf{z}$ representation allows to apply standard multivariate analysis techniques designed for Euclidean vectors to CoDa. For example, one can apply classical Principal Component Analysis to the transformed variables $\boldsymbol{\xi}$ for exploratory data analyses of CoDa. Also, clustering algorithms (i.e. $k-$means) can be applied on the $\boldsymbol{\xi}$-representation of CoDa, without further ado. On the modeling side, any classical multivariate distribution for $\boldsymbol{\xi}$ on $\mathbb{R}^{d-1}$ gives, by back-transformation, a corresponding CoDa distribution for $\mathbf{p} \in \Delta^{d-1}$.

3.7. Description in terms of iterated partitions and some variants

The rescaled stick-breaking can also be described in terms of the amalgamation, subcomposition and partition operations of [2]. Recall that given a CoDa $\mathbf{p} \in \Delta^{d-1}$, an amalgamation of order 1 is a mapping

$$\Delta^{d-1} \ni \mathbf{p} \mapsto \mathbf{t} \in \Delta^1,$$

[4] or, more generally, a strictly monotone continuous function.

obtained when the parts of a $d-$ composition are separated into 2 mutually exclusive and exhaustive subsets, and the composition within each subset are added together. This results in a $2-$parts composition. For example, $\mathbf{p} = (p_1, p_2, p_3, p_4) \in \Delta^3$ can be amalgamated into $\mathbf{t} = (t_1, t_2)$ with $t_1 = p_1 + p_2$, $t_2 = p_3 + p_4$. A subcomposition

$$\Delta^{d-1} \ni \mathbf{p} \mapsto \mathbf{c} \in \Delta^{k-1}$$

is obtained by selecting k parts of a composition and closing the selected subvector to obtain a subcomposition in Δ^{k-1}. Finally, a partition of order 1 is the separation of a $d-$parts composition into two disjoint and exhaustive subsets, and recording the amalgamation and subcomposition of each subsets. For example, the order 1 partition

$$(p_1, \dots, p_k \big| p_{k+1}, \dots, p_d)$$

cuts the $d-$parts at position $1 \leq k \leq d-1$ and yields an amalgamation vector $\mathbf{t} = (t_1, t_2)$, with $t_1 = (p_1 + \dots, p_k)$, $t_2 = (p_{k+1} + \dots + p_d)$, together with the two vectors of subcompositions

$$\mathbf{c}_1 = C(p_1, \dots, p_k) = \frac{(p_1, \dots, p_k)}{t_1}, \quad \mathbf{c}_2 = C(p_{k+1}, \dots, p_d) = \frac{(p_{k+1}, \dots, p_d)}{t_2}.$$

By Property 2.10 and 2.11 of [2], this results in a bijective transformation

$$\Delta^{d-1} \ni \mathbf{p} \mapsto (\mathbf{t}, \mathbf{c}_1, \mathbf{c}_2) \in \Delta^1 \times \Delta^{k-1} \times \Delta^{d-k-1}.$$

The rescaled stick breaking transformation $\mathcal{R}$ of (9) can be described as iterated partitions of order 1 with one subcomposition consisting of a singleton. Indeed, in the particular of partition of order one at position $k = 1$ where $\mathbf{p}$ is partitioned into $(p_1 \big| p_2, \dots, p_d)$, this yields

$$\mathbf{t} = (p_1, 1 - p_1), \quad \mathbf{c}_1 = C(p_1) = 1, \quad \mathbf{c}_2 = C(p_2, \dots, p_d) = \left(\frac{p_2}{1 - p_1}, \dots, \frac{p_d}{1 - p_1} \right)$$

Since $\mathbf{c}_1$ is not informative and $\mathbf{t} \in \Delta^1$, one can only record the first component $t_1 = p_1$ of $\mathbf{t}$ and $\mathbf{c}_2$, i.e. consider the transformation

$$\Delta^{d-1} \ni \mathbf{p} \mapsto (t_1, \mathbf{c}_2) \in [0,1] \times \Delta^{d-2}.$$

In a second stage, one proceeds with another partition of order 1 at position $k = 1$ of the subcomposition $\mathbf{c}_2$ into

$$\left(\frac{p_2}{1 - p_1} \bigg| \frac{p_3}{1 - p_1}, \dots, \frac{p_d}{1 - p_1} \right).$$

This yields the amalgamation vector

$$\mathbf{t}' = \left(\frac{p_2}{1 - p_1}, 1 - \frac{p_2}{1 - p_1} \right) = \left(\frac{p_2}{1 - p_1}, \frac{1 - s_2}{1 - p_1} \right),$$

where $s_2 = p_1 + p_2$, and the two subcompositions vectors

$$\mathbf{c}_1' = 1, \quad \mathbf{c}_2' = C\left(\frac{p_3}{1 - p_1}, \dots, \frac{p_d}{1 - p_1} \right) = \left(\frac{p_3}{1 - s_2}, \dots, \frac{p_d}{1 - s_2} \right).$$

Again, one records the first component of $\mathbf{t}'$, viz. $p_2/(1 - p_1)$, so that one obtains a transformation

$$\Delta^{d-1} \ni \mathbf{p} \mapsto \left(p_1, \frac{p_2}{1 - p_1}, \mathbf{c}_2' \right) \in [0,1]^2 \times \Delta^{d-3}.$$

The reader will readily see that after $d - 1$ iterations one has obtained the $\mathbf{z}$ coordinates which thus correspond to the record of the successive amalgamation vectors in the iterated order one partitions. Note that this gives an alternative proof of Proposition 4 i).

Remark 4. This partition view suggests variants of the rescaled stick-breaking transformation: for example, instead of performing iterated nested partitions from the left to the right at the same position $k = 1$, one can consider a sequence of partitions on the same original CoDa point $\mathbf{p} \in \Delta^{d-1}$, but with varying k from 1 to d. One then records the last component of the first subcomposition of each partition. In other words, one records the closure w.r.t to the left subcomposition of the boxed components $p_1, p_2, \ldots, p_{d-1}$ in the partitions below:

$$\begin{aligned}
&\left(\boxed{p_1}\middle| p_2, \ldots, p_d\right)\\
&(p_1, \boxed{p_2}\,| p_3, \ldots, p_d)\\
&\ldots\\
&\left(p_1, p_2, \ldots, \boxed{p_{d-1}}\middle| p_d\right)
\end{aligned}$$

More explicitly, this amounts to considering the transformation

$$\Delta^{d-1} \ni \mathbf{p} \mapsto \mathbf{z}' = (\mathbf{z}'_1, \ldots, \mathbf{z}'_{d-1}) \in [0,1]^{d-1}$$

defined by

$$\begin{aligned}
\mathbf{z}'_1 &= p_1,\\
\mathbf{z}'_i &= \frac{p_i}{s_i} = \frac{p_i}{p_1 + \ldots + p_i}, \quad i = 2, \ldots, d-1.
\end{aligned}$$

Except for the first term z'_1, this amounts to a $\mathbf{z}$ transform (9), with reverse ordering of the components.

More generally, any transformation whose ith component is a fraction with the numerator being p_i and denominator a sum of p_i and at least another component will lead to a transformed CoDa in the unit cube. For example, one can take

$$\mathbf{z}''_i = \frac{p_i}{p_i + p_{i+1}} \in [0,1], \quad i = 1, \ldots, d-1,$$

with the convention that $0/0 := 0$. However, these variants usually lead to less tractable formulas for the inverse transform and are less prone to interesting statistical interpretations. Hence, we do not pursue further.

4. Application of unit cube geometry: Copulas for CoDa

In the complete neutrality view of [9], each rescaled component z_i does not influence the remaining ones. If one moves out of independence, one can generalize in several directions.

4.1. CoDapulas: copulas for CoDa

As first generalization, one can construct copula models ([34]) for CoDa: instead of taking independent z_i, one can specify a joint distribution for $\mathbf{z}$ by a set of marginals (F_{z_i}), $i = 1, \ldots, d-1$, (each with support the unit interval) and a copula C, viz.

$$\mathbf{z} \sim C(F_{z_1}, \ldots, F_{z_{d-1}}).$$

By back transformation (10), this allows to define general distributions for CoDa points from the specification of a copula and the marginal distributions of the $\mathbf{z}$.

A probabilistic construction of this specification is as follows: let $\mathbf{v} \in [0,1]^{d-1}$ be distributed according to a copula function C, i.e. a multivariate distribution with uniform marginals, and let $Q_{z_i} = F^{-1}_{z_i} : [0,1] \to [0,1]$, $i = 1, \ldots, d-1$, be univariate quantile functions with range $[0,1]$. Set $z_i = Q_{z_i}(v_i)$, $i = 1, \ldots, d-1$. Then $\mathbf{z} = (z_1, \ldots, z_{d-1}) \in [0,1]^{d-1}$

has copula function C and marginal distributions (F_{z_i}). By back transformation, $\mathbf{p} \in \Delta^{d-1}$ is a CoDa point whose distribution is uniquely specified by C an the set (F_{z_i}) of marginal cdfs. Explicitly, (10) yields

$$p_1 = Q_{z_1}(v_1), \tag{16}$$

$$p_i = Q_{z_i}(v_i)\prod_{j=1}^{i-1}(1 - Q_{z_j}(v_j)), \quad i = 2, \ldots, d-1, \tag{17}$$

$$p_d = \prod_{i=1}^{d-1}(1 - Q_{z_i}(v_i)). \tag{18}$$

Conversely, given a CoDa point $\mathbf{p}$, one can estimate and study its intra-dependence through the copula of its $\mathbf{z}$-representation: one first transforms $\mathbf{p}$ into $\mathbf{z}$ by transformation (9), and then standardize the marginals z_i to the uniform distribution. The latter operation is obtained, when $\mathbf{z}$ is continuous, by the marginal probability integral transforms,

$$v_i := F_{z_i}(z_i), \quad i = 1, \ldots, d-1,$$

where F_{z_i} is the c.d.f. of z_i. Then, $\mathbf{v} = (v_1, \ldots, v_{d-1})$ has uniform marginals, i.e. has a copula distribution. See Figure 7.

$$\begin{tikzcd}
\mathbf{p} \in \Delta^{d-1} \arrow[r, bend left] & \mathbf{z} \in [0,1]^{d-1} \arrow[l, bend left] \arrow[r, bend left, "F_{z_i}"] & \mathbf{v} \sim C \arrow[l, bend left, "Q_{z_i}"]
\end{tikzcd}$$

Fig. 7: Probabilistic construction of a CoDapula from its rescaled stick breaking representation.

(In the non-continuous case, the standardization is obtained by using the marginal distributional transforms instead. The latter is defined as

$$F_{z_i}(x, \eta) := P(z_i < x) + \eta P(z_i = x), \quad \eta \in [0,1]. \tag{19}$$

Then, $\mathbf{v}$ is obtained by setting

$$v_i := F_{z_i}(z_i, \eta_i), \quad i = 1, \ldots, d-1,$$

where (η_i) is a sequence of i.i.d. uniformly distributed on $[0,1]$ randomizers, independent of $\mathbf{z}$, see e.g. [42], [16]).

Let us give a fancy name to the copula of a CoDa point.

Definition 5 (CoDapula)**.** Let $\mathbf{p} \in \Delta^{d-1}$ a random CoDa point, and $\mathbf{z} \in [0,1]^{d-1}$ be it rescaled stick-breaking representation (9). Then, a CoDapula of (the distribution of) $\mathbf{p}$ is a copula of (the distribution of) $\mathbf{z}$. In other words, a CoDapula C of $\mathbf{p}$ is the distribution of $\mathbf{v}$ in the construction of Figure 7.

Thanks to Sklar's Theorem ([44]), a CoDapula always exists. It is unique if $\mathbf{z}$ is continuous (see e.g. [34]). Definition 5 depends on the ordering of the components $1, \ldots, d$. Hence, a CoDapula of $\mathbf{p}$ depends on a permutation π of $\{1, \ldots, d\}$. Hence, in full rigor, one should have defined a notion of $\pi-$CoDapula to stress the dependence on π. We have chosen not to in order to simplify notations. The choice of the ordering, i.e. of π, may depend on the application in view, and will be discussed in Section 6.2.

By (16), the first component p_1 has the same distribution as z_1, and thus is completely specified by the first marginal distribution function F_{z_1}, (equivalently, quantile function Q_{z_1}). Note that the marginal distributions of the remaining components $p_2, \ldots, p_d$ depend on both the CoDapula and the marginal distributions: this is in contrast with the copula approach for classical Euclidean vectors. Nonetheless, at the $\mathbf{z}$ level, one has the the classical copula separation of a multivariate distribution into its marginal distributions $F_{z_1}, \ldots F_{z_{d-1}}$ and the dependence structure embodied in the copula function C.

4.2. Examples and numerical illustrations

We illustrate in Figures 8 and 9 some CoDa distributions which can be obtained using the specification by a CoDapula of Definition 5 and marginal quantile functions for **z**. In Figure 8, the copula of **z** is an Ali-Mikhail-Haq copula with parameter $\alpha = 0.91$, and the marginal distributions are Beta and uniform, $F_{z_1} \sim \beta(1/4, 2)$ and $F_{z_2} \sim U_{[0,1]}$, while in Figure 9, the CoDapula is a Gumbel-Hougaard copula with parameter $\theta = 7$, and same marginals as in Figure 8. The left panels show scatterplots in the **z** domains, and the right panels the corresponding ternary plots in the simplex CoDa space for **p**. Figure 8 give an example of mild dependence with CoDa points spreading above the $p_1 = 0$ level. As the Gumbel copula approaches the comonotonicity copula as $\theta \to \infty$ (while $\theta = 1$ yields the independence copula), the value $\theta = 7$ in Figure 9 models strong dependence at the **z** level, resulting in the pattern of points shown on the right panel at the CoDa level. More examples could be considered.

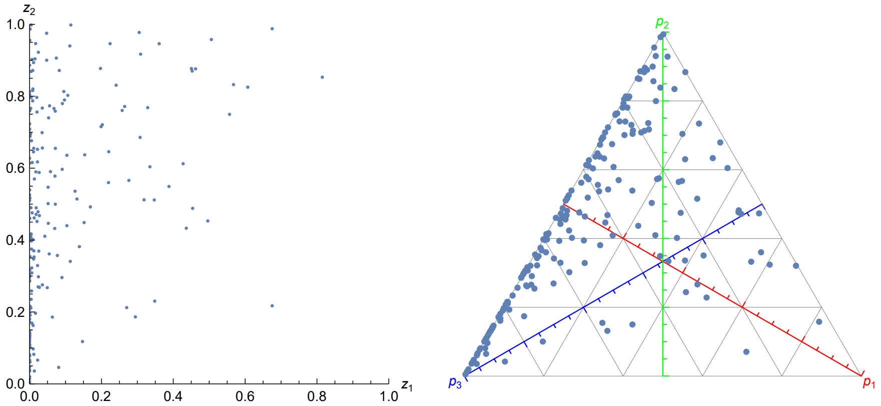

Fig. 8: CoDa with AMH CoDapula ($\alpha = 0.91$), and $\beta(1/4, 2)$, $U_{[0,1]}$ marginal distribution functions: scatter plot at the **z** level (left) and ternary scatter plot for the resulting **p** (right), $d = 3$.

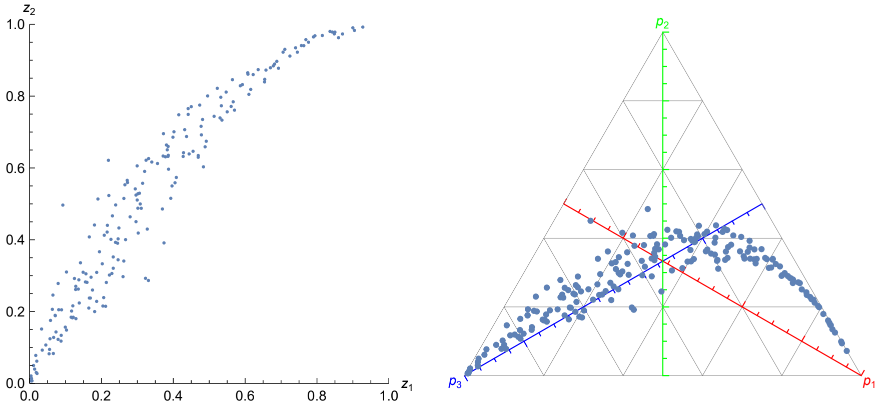

Fig. 9: CoDa with Gumbel CoDapula ($\rho = 7$) and $\beta(1/4, 2)$, $U_{[0,1]}$ marginal distribution functions; scatter plot at the **z** level (left) and ternary scatter plot for the resulting **p** (right), $d = 3$.

4.3. Complete dependence of CoDa

The concept of CoDapula opens the gates of the vast copula literature and modeling methodology to CoDa. This is useful to study the intra-dependence of CoDa. The independence copula for **z** means that **p** is completely neutral. At another extreme, complete dependence at the z level induces a specific dependence pattern at the **p** level, as is shown in the next two examples.

4.3.1. Comonotone CoDapula

Comonotonicity is an extreme form of dependence structure for Euclidean vectors that describes the strongest positive dependence. A comonotone vector is characterised by having as copula the comonotone copula $M(x_1, \dots, x_{d-1}) = \min(x_1, \dots, x_{d-1})$. The comonotone copula corresponds to the distribution of the vector $\mathbf{v} = (v, \dots, v) \in [0, 1]^{d-1}$, with a single $v \sim U_{[0,1}$. In other words,

$$P(\mathbf{v} \le \mathbf{x}) = P(v \le x_1, \dots, v \le x_{d-1}) = \min(x_1, \dots, x_{d-1}), \quad \mathbf{x} \in [0, 1]^{d-1}.$$

Applied to CoDa, the corresponding $\mathbf{z}$ thus writes

$$\mathbf{z} = (Q_{z_1}(v), \dots, Q_{z_{d-1}}(v)),$$

where $Q_{z_i} : [0, 1] \to [0, 1]$ are given quantile functions. This gives, as corresponding $\mathbf{p}$, Coda with components

$$\begin{aligned} p_1 &= Q_{z_1}(v) \\ p_i &= Q_{z_i}(v) \prod_{j=1}^{i-1} \left(1 - Q_{z_j}(v)\right), \quad i = 2, \dots, d-1. \end{aligned}$$

Example 4 (Comonotone CoDapula, $d = 3$)**.** *For example, for $d = 3$, one gets*

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} Q_{z_1}(v) \\ Q_{z_2}(v) \end{pmatrix}$$

which translates into

$$\begin{aligned} p_1 &= Q_{z_1}(v) \\ p_2 &= Q_{z_2}(v)\,(1 - Q_{z_1}(v)) \\ p_3 &= 1 - p_1 - p_2 = (1 - Q_{z_1}(v))\,(1 - Q_{z_2}(v)) \end{aligned} \tag{20}$$

Thus, p_1 is an increasing function of v, p_3 is decreasing, while p_2 switches direction of variation w.r.t v.

Figure 10 shows, for $d = 3$, the CoDa $\mathbf{p}$ corresponding to the comonotone copula, with uniform quantile functions at the $\mathbf{z}$ level, viz. $Q_{z_1}(v) = Q_{z_2}(v) = v$ in (20), so that $p_1 = v$, $p_2 = v(1 - v)$, $p_3 = (1 - v)^2$.

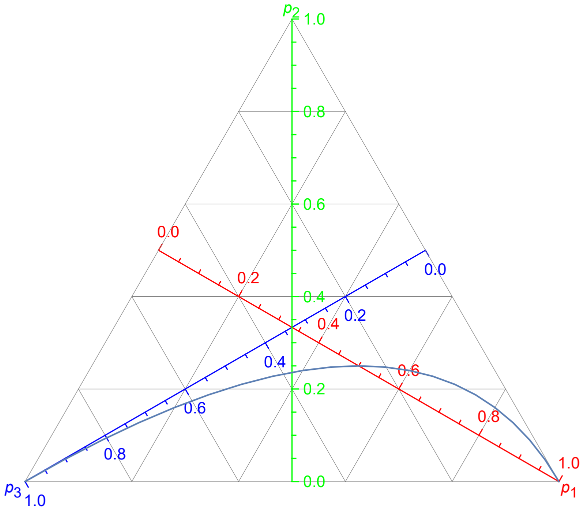

Fig. 10: Ternary plot of a CoDa with comonotone CoDapula and uniform quantile functions, $d = 3$, with barycentric axes p_1 (red), p_2 (green), p_3 (blue).

The distribution of **p** *is singular, as each component p_i is a deterministic function of $v \sim U_{[0,1]}$:* **p** *lies on the curve shown in the ternary plot. This implies that each pair of components (p_i, p_j), $1 \leq i \neq j \leq 3$ are totally dependent, i.e. lie on a curve. Figure 11 shows the resulting complete dependence between each pairs of components: (p_1, p_3) are counter-monotone (middle), while (p_2, p_3) (right) is comonotone. (p_1, p_2) (left) switches its sense of variation, being first comonotone, then countermonotone. Note that p_2 also has a limited range of variation $p_2 \in [0, 1/4]$.*

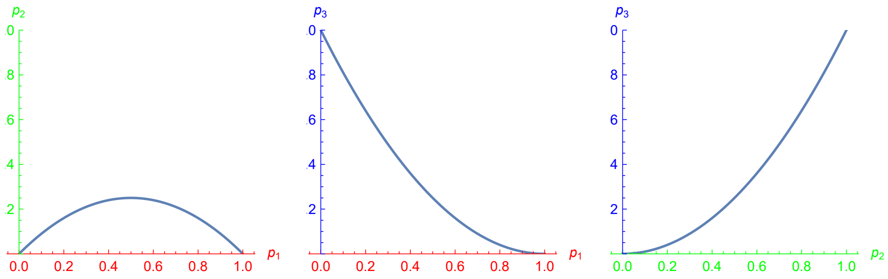

Fig. 11: Complete dependence between pairs of components of a CoDa with comonotone CoDapula and uniform quantile functions, $d = 3$: (p_1, p_2) (left), counter-monotone (p_1, p_3) (middle), comonotone (p_2, p_3) (right).

4.3.2. Counter-monotone CoDapula

Counter-monotonicity is the antithesis of comonotonicity. Note that this notion is well-defined only in two dimensions. We thus restrict our discussion to the case $d = 3$. The bivariate counter-monotone copula $W(x_1, x_2) = \max(x_1 + x_2 - 1, 0)$ is stochastically realized by the vector $\mathbf{v} = (v, 1 - v)$, where $v \sim U_{[0,1]}$. This gives a corresponding Coda

$$\mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} Q_{z_1}(v) \\ Q_{z_2}(1-v)\,(1 - Q_{z_1}(v)) \\ (1 - Q_{z_1}(v))\,(1 - Q_{z_2}(1-v)) \end{pmatrix}. \tag{21}$$

The following example illustrates the case when the quantile functions are the uniform ones.

Example 5 (Counter-monotone CoDapula)**.** *For $Q_{z_1}(v) = Q_{z_2}(v) = v$, (21) gives $p_1 = v$, $p_2 = (1 - v)^2$, $p_3 = v(1 - v)$. One thus gets the same parametrization at the CoDa level as in the comonotone case of Example 4, but with the roles of p_2 and p_3 exchanged, see Figure 12.*

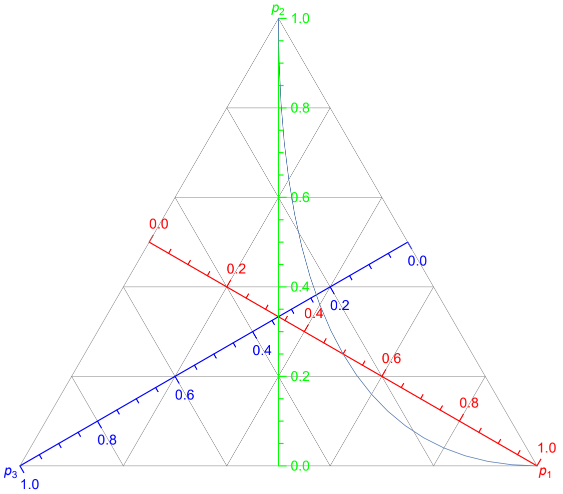

Fig. 12: Ternary plot of a CoDa with countermonotone CoDapula and uniform quantile functions, $d = 3$, with barycentric axes p_1 (red), p_2 (green), p_3 (blue).

This now translates at the CoDa level into complete dependence between pairs of components, as shown in Figure 13. Notice, however, that the dependence pattern is not the symmetric of the comonotone case of Figure 11: (p_1, p_2) are now counter-monotone, whereas both $(p_1, p_3$ and (p_2, p_3) change their direction of variation. Hence, with a counter-monotone CoDapula, only one pair is monotone dependent (viz. (p_1, p_2) counter-monotone), whereas with a comonotone CoDapula, two pairs were monotone dependent (viz. (p_1, p_3) counter-monotone, and (p_2, p_3) comonotone).

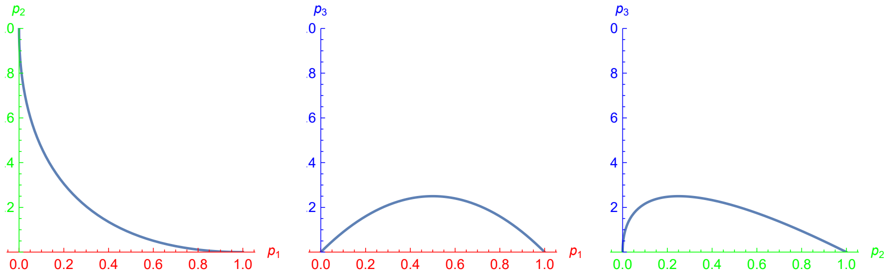

Fig. 13: Complete dependence between pairs of components of a CoDa with countermonotone CoDapula and uniform quantile functions, $d = 3$: counter-monotone (p_1, p_2) (left), (p_1, p_3) (middle), (p_2, p_3) (right)

5. Application of unit cube geometry: Regression models for CoDa

As a second possible generalization, the rescaled stick-breaking approach (9) can be useful for the intra regression analysis of a CoDa component w.r.t the others. The basic idea is to construct regression models in the $\mathbf{z}$ coordinates to iteratively explain one z_i component in terms of the other z_j. Indeed, the transformation (9) is reminiscent of Rosenblatt's generalization of the quantile transform by successive conditioning and the regression representation of a random vector, which we recall now.

5.1. Regression representation of an Euclidean random vector

Let $\mathbf{X} = (X_1, \dots X_k) \in \mathbb{R}^k$ be a vector with joint c.d.f. F. If one can transform $\mathbf{X}$ into a sequence $\epsilon_1, \dots, \epsilon_k$ of independent, identically distributed r.v., with a prescribed distribution λ (say, uniform on [0, 1]), then, one can argue that the distribution of $\mathbf{X}$ has been successfully modeled: the transformation

$$(X_1, \dots, X_k) \xrightarrow{\phi} (\epsilon_1, \dots, \epsilon_k)$$

has stripped $\mathbf{X}$ of all its stochastic variability and dependence and turned it into white noise. The function ϕ effectively models the distribution F of $\mathbf{X}$.

[41]'s transform and its generalizations (see [42]) achieves such a reduction: Denote by $F_{i|i-1,\dots,1}$ the conditional c.d.f. of X_i given $(X_{i-1}, \dots, X_1)$, $i = 2, \dots, k$, with F_1 the (marginal) cdf of X_1. [41]'s transform is then defined by $\boldsymbol{\epsilon} = (\epsilon_1, \dots \epsilon_k)$ with

$$\begin{aligned} \epsilon_1 &:= F_1(X_1) \\ \epsilon_i &:= F_{i|i-1,\dots,1}(X_i | X_{i-1}, \dots, X_1), \quad i = 2, \dots, k. \end{aligned}$$

Under an assumption of continuity[5] of the successive conditional c.d.f. $F_{i|i-1,\dots,1}$, Rosenblatt's transform turn the vector $\mathbf{X}$ into a vector $\boldsymbol{\epsilon}$ of i.i.d. $U_{[0,1]}$ components (see [42]).

Conversely, starting from a vector $\boldsymbol{\epsilon} \sim \lambda^k$ and applying the successive (conditional) quantile functions $F^{-1}_{i|i-1,\dots,1}$, viz.

$$\begin{aligned} X_1 &:= F_1^{-1}(\epsilon_1) \\ X_i &:= F^{-1}_{i|i-1,\dots,1}(\epsilon_i | X_{i-1}, \dots, X_1), \quad i = 2, \dots, k, \end{aligned} \tag{22}$$

[5] For discontinuous conditional cdf, one must use the conditional probability integral transform (19) instead, see [42]

one obtains a vector $\mathbf{X}$ with the desired joint c.d.f. F. Each equation (22) interprets as a nonlinear regression equation of X_i, given its past covariates X_j, $j < i$, with error/noise/innovation ϵ_i. This gives a regression representation of $\mathbf{X}$, according to [42], where (22) is a (triangular) stochastic representation of the successive predictive distributions $P^{X_i|X_{i-1},\ldots,X_1}$.

In (22), the distribution of $(\epsilon_1, \ldots, \epsilon_d)$ is purely conventional, the only constraint is that it be absolutely continuous (so that any distribution of $\mathbf{X}$ can be obtained from it by mapping and not by Markov kernels, see [16]). In particular, one can choose the more familiar Gaussian white noise framework by setting

$$\epsilon_i = \phi(\epsilon_i'),$$

where $\epsilon_1', \ldots, \epsilon_k'$ are i.i.d. standard Gaussian $\mathcal{N}(0, 1)$, and ϕ is the c.d.f. of the $\mathcal{N}(0, 1)$ distribution.

The regression representation (22) is the general, exact, nonlinear form of a regression model. In particular, if the conditional quantile functions $F^{-1}_{i|i-1,\ldots,1}$ are linear, one obtains the classical linear model (albeit with uniform noise), viz.

$$X_i = a_{i,1}X_1 + \ldots + a_{i,i-1}X_{i-1} + \epsilon_i,$$

where $a_{i,1}, \ldots, a_{i,i-1}$ are parameters.

5.2. *Parametric internal regression models for CoDa*

This suggests to make use of this regression representation to construct triangular regression models on the z representation of CoDa, by applying the transformation (22) to the z_i of (9) instead of the X_i: each z_i is explained in terms of the previous z_j, $j < i$, and some extraneous randomness $\epsilon_i \sim \lambda$, for $i = 1, \ldots, d-1$. Then, by back transformation (9), one obtains a (possibly nonlinear) regression model for the original p_i, which can be used for internal prediction of a component in term of the others.

More precisely, let $\boldsymbol{\epsilon} \sim \lambda^{d-1}$ be a vector of uniform noise on $[0, 1]^{d-1}$. Then, a general nonlinear triangular regression model for the z writes

$$\begin{aligned} z_1 &= \phi_1(\epsilon_1) \\ z_i &= \phi_i(\epsilon_i, z_{i-1}, \ldots, z_1), \quad i = 2, \ldots, d-1, \end{aligned} \tag{23}$$

where $\phi_i : [0, 1]^i \mapsto [0, 1]$ are s.t. $\epsilon_i \to \phi_i(\epsilon_i, z_{i-1}, \ldots, z_1)$ is non-decreasing, left-continuous, with $\phi_i(0, z_{i-1}, \ldots, z_1) = 0$, $\phi_i(1, z_{i-1}, \ldots, z_1) = 1$. (i.e. the ϕ_i satisfy the properties of univariate quantile functions).

For example, a Gausssian (partially)linear triangular model can be obtained by specifying the error distribution as standard multivariate Gaussian $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I_{d-1})$, and the z_i as

$$\begin{aligned} z_1 &= \Phi(\epsilon_1) \\ z_i &= \Phi(a_{i,1}z_1 + \ldots + a_{i,i-1}z_{i-1} + \epsilon_i), \quad i = 2, \ldots, d-1 \end{aligned} \tag{24}$$

where Φ, the cdf of the standard univariate Gaussian distribution, is applied to ensure that $z_i \in [0, 1]$. More generals models can be constructed via Generalized Linear Models, see e.g. [33], and more specifically, [6] for data on the unit interval $[0, 1]$.

5.3. *Example: agriculture data.*

We provide below a basic example of the construction of a parametric internal regression model in the $\mathbf{z}$ space for CoDa, illustrated on a real dataset. The data is taken from the example datasets accompanying Mathematica's ([52]) "TernaryListPlot" command. It gives the raw amount of fertilizers (Nitrogen-Potassium-Phosphate) in a time series from 1960 to 2015. The scatter plots, both at the $\mathbf{z}$ level (left panel), and at the compositional level in the ternary plot (right panel) in Figure 14 show a cyclic pattern in the composition of fertilizers. The sinusoidal shape of the transformed data at the $\mathbf{z}$ level (left panel) suggests the following model,

$$z_2 = a_0 + a_1 \cos(\omega z_1 + \phi) + \epsilon,$$

where a_0, a_1, ω, ϕ are parameters to be estimated and ϵ the random error. Nonlinear least squares (command "NonLinearModelFit" in [52]) gives as fitted model,

$$z_2 \approx 0.5 + 0.047 \cos(-21.86\, z_1 + 8.76),$$

with sums of squares of 13.43 for the model and 0.0075 for the error. In terms of the original CoDa variables $\mathbf{p}$, this yields a predictive model of the components in term of the first one p_1,

$$\begin{aligned} p_1 &= p_1 \\ p_2 &\approx (1 - p_1)(0.5 + 0.047 \cos(-21.86\, p_1 + 8.76)) \\ p_3 &\approx 1 - p_1 - p_2 \approx (1 - p_1)(0.5 - 0.047 \cos(-21.86\, p_1 + 8.76)). \end{aligned}$$

The resulting fitted curve is shown in orange in Figure 14, with extrapolated values on the full range $z_1 = p_1 \in [0, 1]$.

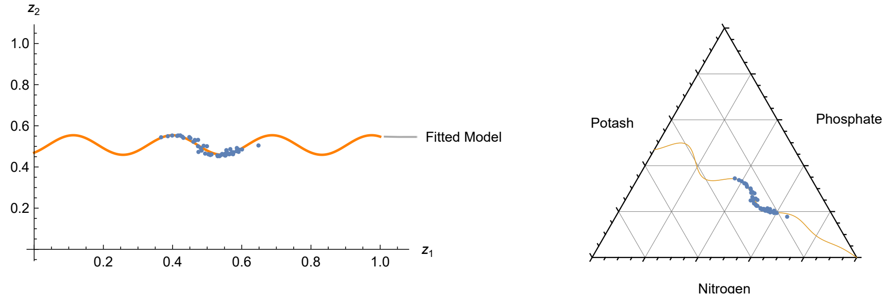

Fig. 14: Internal parametric regression model for agriculture data. Scatter plots (blue points) and fitted sinusoidal model (orange line) in the $\mathbf{z}$ space (left) and for the original data (right).

5.4. Extensions and alternatives

As shown in the previous example of Section 5.3, the rescaled stick-breaking transformation (9) reduces internal regression analysis of CoDa to classical regression analysis of vector data. Hence, all classical multivariate regression analysis techniques apply to CoDa, in their transformed $\mathbf{z}$ representation of the free unit cube. For space constraints, we limited ourselves in the example of Section 5.3 to a very basic illustration with a parametric regression model. Let us thus briefly mention some extensions and alternatives:

- A nonparametric alternative to the above intra-parametric models is to directly start from (22) and estimate the conditional distributions of z_i given $(z_{i-1}, \ldots, z_1)$, or some functional thereof, via some nonparametric estimate. For example, one can look for the mean of these conditional distributions, and estimate the regression function of $E[z_i | z_{i-1}, \ldots, z_1]$ by a Nadara-Watson, spline, or local polynomial estimator.

- Many applications are interested in explaining/predicting a CoDa point $\mathbf{p}$ w.r.t. some covariates $\mathbf{X} \in \mathbb{R}^k$, i.e. in studying the conditional distribution of $\mathbf{p}|\mathbf{X} = \mathbf{x}$. This can be done via the rescaled $\mathbf{z}$ representation (9) by performing a regression of $\mathbf{z}$ w.r.t. the covariates $\mathbf{X}$. As said before, one must ensure that the constraint $\mathbf{0} \leq \boldsymbol{z} \leq \mathbf{1}$ is fulfilled. This can be achieved by a link function which entails the correct normalisation or by mapping $\mathbf{z} \in [0, 1]^{d-1}$ into $\boldsymbol{\xi} \in \mathbb{R}^{d-1}$, by using the device explained in Section 3.6. In parametric regression models, one can incorporate these external covariates $\mathbf{X}$ by making the parameters of $\mathbf{p}$ in its rescaled $\mathbf{z}$ representation (9), (i.e. the functions ϕ_i in (23) or the coefficients $a_{j,i}$ in (24)) as functions of the covariates $\mathbf{X}$. Once a regression model/ or a nonparametric estimate for $\mathbf{z}$ given $\mathbf{X}$ has been computed, one back transforms the predicted values $\hat{\mathbf{z}}$ into predicted values of $\hat{\mathbf{p}}$, via the inverse transformation (10).

- Our focus in this paper is on internal dependence analysis of CoDa. However, one can also easily envision external regression analysis such as CoDa to CoDa or CoDa to vector, by conducting a similar analysis in the corresponding $\mathbf{z}$ space for the CoDa input/output variables considered.

- In addition to explicit regression models, one can also assess quantitatively neutrality of z_1 by the strength of the regression dependence between z_1 and the remaining z_j components of p, for $j > 1$. This quantification can be achieved through multivariate asymmetric correlation coefficients, like the recent [25]'s $\zeta^1(\mathbf{X}, Y)$ or [4]'s $T(\mathbf{X}, Y)$[6], the latter being a multivariate extension of the bivariate measure ξ of [8]. These coefficients quantifies the extent of regression dependence of a univariate random variable Y on a k-dimensional random vector $\mathbf{X} = (X_1, \dots, X_k)$: they are equal to 0 in case of independence, and equal to 1 if Y is measurable function of $\mathbf{X}$. Applied to our context, one can thus quantify the amount of (non-)neutrality of p_1 by computing $\zeta^1((z_2, \dots, z_d), z_1)$ or $T((z_2, \dots, z_d), z_1)$.

6. Conclusion and further remarks

6.1. Conclusion

We have thus brought to the fore these two related transformations for CoDa based on stick-breaking processes. The first one represents a CoDa point as a set of ordered values on the unit interval, whereas the second one, which originates from [27] and [9], removes the unit-sum constraint of the simplex representation and turns a CoDa point into a free vector of the unit cube. It is noticeable that these stick-breaking representations are not mentioned in the reference books [2], [23], [36], [19], [7] on CoDa and thus do not seem to be well-known inside the CoDa community. We have thought it was thus commendable to publicize them and consolidate the various connections they have with several topics of probability and statistics into a cohesive account.

Both approaches are useful to construct distributions for CoDa from multivariate distributions of Euclidean vectors. The second approach appears most promising as it allows for a reduction of CoDa points to classical multivariate vectors and thus allows the use of well-established multivariate analysis techniques and models to be directly transferred to CoDa. Important statistical models and tools like generalised linear models, graphical models, vines/factor copulas, clustering, Principal Component Analysis, non-parametric and semi-parametric techniques, etc. are now at the disposal of the Statistician and beg for their application to CoDa. In particular, we are confident that the concept of a CoDapula, a copula for CoDa, is promising as it allows to study the intra-dependence of CoDa with such copulas.

Let us stress that these stick-breaking representations allow to deal effectively and simply with CoDa with zeroes: in the ordered representation (3), zeroes translates into ties in $\mathbf{s}$, while in the rescaled $\mathbf{z}$ representation, zeroes of $\mathbf{p}$ translates into $\mathbf{z}$ being sent to the boundary of $[0, 1]^{d-1}$. Distributionally, this means that the (u_i) in Definition 2 of the ordered representation (3) have common discrete components in their distributions. For the rescaled $\mathbf{z}$ representation (9), zeroes of $\mathbf{p}$ translates distributionally into $\mathbf{z}$ having a singular component on some faces of $[0, 1]^{d-1}$ in the Lebesgue decomposition of the probability measure of $\mathbf{z}$. One can therefore use mixed/general distributions to model such CoDa points with possibly zero components.

We thus believe these stick-breaking representations provide an interesting complementary approach to the classical log-ratio coordonatizations techniques of Aitchison and his followers. For length reasons, we have barely scratched the surface of statistical applications based on these transformations. Much more needs to be done to explore its potentialities and eventual limitations. The multi-faceted aspect of CoDa, which can be envisioned from so many viewpoints, via the log-ratios, the stick-breaking, the projective geometry ([15]), or the manifold and information geometric ([13]) approaches, is what makes CoDa such a fascinating topic.

Let us close the article with some further remarks.

6.2. Choice of the ordering of the components

A possible issue of the transformations (2) and (9) is the lack of symmetry w.r.t. the components, as they depend on the ordering of the components $1, \dots, d$ of the composition. The question thus arises which ordering is most adequate. Several possibilities can be envisioned.

[6] Note that [4] introduce a more general regression dependence coefficient which allows for covariates and the assessment of *conditional* independence.

- A first possibility is to let the Statistician decide for himself. This is similar to the "working-in-coordinates" principle in classical log-ratio CoDa analysis: the statistical model is extrinsic and built w.r.t. a given coordinate frame (here, the ordering chosen), and is eventually mapped back to the original simplex. This was the approach chosen in the example of Section 5.3: the sinusoidal regression model in the $\mathbf{z}$ space is mapped to the CoDa simplex space and then gives a model which explains/predicts how the remaining components p_2, p_3 are driven by p_1.

- The order of the components can be dictated by the type of application in view. For example, in a general regression model $Y = r(X, \epsilon)$, there is a natural asymmetry in the vector (X, Y) between the dependent/predicted variable Y and the independent/predictor variables X: one wants to explain/predict Y *from* X (with noise ϵ). There is also an asymmetry in the rescaled stick-breaking transformation (10), between the first component p_1 and the remaining ones $p_2, \ldots, p_d$: the first component is identical in the simplex space $\mathbf{p}$ representation as in the unit cube $\mathbf{z}$ representation, i.e. $p_1 = z_1$, whereas the remaining components $p_2, \ldots, p_d$ depend on several of the z_i. (p_i is a function of z_j, for $1 \leq j \leq i$). Thus, z_1 is directly interpretable as one original component and a statistical analysis of z_1 translates into a statistical analysis of the first component p_1. So, if one is interested in evaluating how a specific component is influenced by the remaining parts, it is sensible to take this component as first one: a regression model $z_1 = r(z_2, \ldots, z_d, \epsilon)$ at the $\mathbf{z}$ level, following the methodology explained in Section 5, with $z_1 = p_1$ as predicted variable, directly gives a regression model of the first component p_1 in terms of the remaining (rescaled) components, viz. $p_1 = r(z_2, \ldots, z_d, \epsilon)$.

- One can also envision a data-dependent choice of the ordering of the components: the basic idea of the $\mathbf{z}$ transformation is to transform the study of the non-neutrality of the constrained components p_i of the composition into a the study of the dependence of the free z_i. Thus, it would make sense to order the components by decreasing order of non-neutrality/dependence with the remaining composition. If, w.l.o.g, the first component p_1 is most dependent with the remaining composition $(\frac{p_2}{1-p_1}, \ldots, \frac{p_d}{1-p_1})$, it means that p_1 is the main factor explaining the remaining composition. Having isolated such a component, one can then look within the closed remaining composition $(\frac{p_2}{1-p_1}, \ldots, \frac{p_d}{1-p_1})$ of size $d-1$, which component is most dependent with the closed subcompositions of size $d-2$. The process is then iterated, yielding an ordering of the components. In practice, such evaluation of the dependence between a component and a subcomposition can be performed using the estimators of the asymmetric regression dependence coefficients $\zeta^1(\mathbf{X}, Y)$ of [25] or $T(\mathbf{X}, Y)$ of [4], mentioned in Section 5.4.

  This gives the following algorithm: For a composition $\mathbf{p}$ of size d,

  1. Select j s.t. $T(\mathbf{W}_j, p_j)$ (or $\zeta^1(\mathbf{W}_j, p_j)$) is maximum, where $\mathbf{W}_j := \left(\ldots, \frac{p_k}{1-p_j}, \ldots\right)_{k \neq j}$ is the closed subcomposition of size $d-1$ with component j omitted.
  2. Define a new composition $\mathbf{p}' = (p'_k)$ of size $d' = d-1$, with, for $1 \leq k \leq d$, $k \neq j$, $p'_k = p_k/(1-p_j)$, so that $\mathbf{p}'$ has component j removed.
  3. If $d' > 1$, return to step 1, with $\mathbf{p}'$ in lieu of $\mathbf{p}$, and d' instead of d.

- One can also mix the above approaches, e.g. select as first component the one the Statistician is interested in explaining/predicting, and select the remaining ones in a data-dependent manner.

6.3. Connection with mixability: existence of CoDa distributions with given marginals

The stick-breaking approaches were constructive and gave explicit representations of distributions of CoDa points. In case of the ordered approach, the stick-ordered distributions of Definition 2, were parametrized by $d-1$ univariate marginals. Similarly, in the rescaled approach, the distributions are parametrized by either a $d-1$ dimensional copula and $d-1$ marginals, or a set of $d-1$ conditional distributions. These approaches were helpful in constructing distributions for $d-$dimensional CoDa points.

A converse issue is to enquire for the existence of a d-dimensional CoDa distribution with a given set of d marginal distributions. This question is related to the notion of joint mixability, which is a notion mainly investigated in the risk theory literature (See the survey by [49]). This connection between mixability and distributions for CoDa does not seem to have been made beforehand by the CoDa community.

Recall that the definition of joint mixability ([48]) is as follows:

Definition 6. An d-tuple of probability distributions on $\mathbb{R}$, $(F_1, \ldots, F_d)$ is jointly mixable if there exist d random variables $X_1 \sim F_1, \ldots, X_d \sim F_d$ such that $X_1 + \ldots + X_d =: K$ is almost surely a constant.

Hence, the question of existence of a d-dimensional CoDa distribution with given marginals is a special case of mixability with $K = 1$. [20] Theorem 5 give a necessary and sufficient condition. Necessary conditions are given in Theorem 2.1 in [48], and sufficient conditions are given in Theorems 3.1, 3.2, and 3.4 for uniform, monotone, and symmetric-unimodal densities, respectively.

Acknowledgements

Olivier P. Faugeras acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

References

References

[1] J. Aitchison, The statistical analysis of compositional data, Journal of the Royal Statistical Society: Series B (Methodological) 44 (1982) 139–160.

[2] J. Aitchison, The statistical analysis of compositional data, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1986.

[3] T. W. Anderson, An introduction to multivariate statistical analysis, Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

[4] M. Azadkia, S. Chatterjee, A simple measure of conditional dependence, Ann. Statist. 49 (2021) 3070–3102.

[5] R. C. e. Balakrishnan N., Handbook of statistics 16. Order statistics: theory and methods, volume 16, Elsevier Science, 1998.

[6] W. H. Bonat, P. Ribeiro Jr, W. M. Zeviani, Regression models with responses on the unity interval: Specification, estimation and comparison, Biometric Brazilian Journal 30 (2012) 415–431.

[7] K. G. van den Boogaart, R. Tolosana-Delgado, Analyzing compositional data with R, Use R!, Springer, Heidelberg, 2013.

[8] S. Chatterjee, A new coefficient of correlation, Journal of the American Statistical Association 0 (2020) 1–21.

[9] R. J. Connor, J. E. Mosimann, Concepts of independence for proportions with a generalization of the dirichlet distribution, Journal of the American Statistical Association 64 (1969) 194–206.

[10] H. A. David, H. N. Nagaraja, Order statistics, Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

[11] J. D. Dollard, C. N. Friedman, Product integration with applications to differential equations, volume 10 of Encyclopedia of Mathematics and its Applications, Addison-Wesley Publishing Co., Reading, MA, 1979. With a foreword by Felix E. Browder, With an appendix by P. R. Masani.

[12] S. Engen, A note on the geometric series as a species frequency model, Biometrika 62 (1975) 697–699.

[13] I. Erb, N. Ay, The information-geometric perspective of compositional data analysis, in: Advances in compositional data analysis—Festschrift in honour of Vera Pawlowsky-Glahn, Springer, Cham, [2021] ©2021, pp. 21–43.

[14] W. J. Ewens, Population genetics theory—the past and the future, in: Mathematical and statistical developments of evolutionary theory (Montreal, PQ, 1987), volume 299 of NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci., Kluwer Acad. Publ., Dordrecht, 1990, pp. 177–227.

[15] O. P. Faugeras, An invitation to intrinsic compositional data analysis using projective geometry and hilbert's metric, TSE Working Paper, no. 23-1496 (2023).

[16] O. P. Faugeras, L. Rüschendorf, Markov morphisms: a combined copula and mass transportation approach to multivariate quantiles, Math. Appl. (Warsaw) 45 (2017) 21–63.

[17] S. Feng, A note on residual allocation models, Front. Math. China 16 (2021) 381–394.

[18] T. S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Statist. 1 (1973) 209–230.

[19] P. Filzmoser, K. Hron, M. Templ, Applied compositional data analysis, Springer Series in Statistics, Springer, Cham, 2018. With worked examples in R.

[20] N. Gaffke, L. Rüschendorf, On a class of extremal problems in statistics, Math. Operationsforsch. Statist. Ser. Optim. 12 (1981) 123–135.

[21] S. Ghosal, A. van der Vaart, Fundamentals of nonparametric Bayesian inference, volume 44 of Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2017.

[22] R. D. Gill, S. r. Johansen, A survey of product-integration with a view toward application in survival analysis, Ann. Statist. 18 (1990) 1501–1555.

[23] M. Greenacre, Compositional data analysis in practice, Chapman and Hall/CRC, 2018.

[24] M. Greenacre, Compositional data analysis, Annu. Rev. Stat. Appl. 8 (2021) 271–299.

[25] F. Griessenberger, R. R. Junker, W. Trutschnig, On a multivariate copula-based dependence measure and its estimation, Electron. J. Stat. 16 (2022) 2206–2251.

[26] R. C. Griffiths, Exact sampling distributions from the infinite neutral alleles model, Adv. in Appl. Probab. 11 (1979) 326–354.

[27] P. R. Halmos, Random alms, Ann. Math. Statistics 15 (1944) 182–189.

[28] H. Ishwaran, L. F. James, Gibbs sampling methods for stick-breaking priors, J. Amer. Statist. Assoc. 96 (2001) 161–173.

[29] P. Jaworski, T. Rychlik, On distributions of order statistics for absolutely continuous copulas with applications to reliability, Kybernetika (Prague) 44 (2008) 757–776.

[30] M. C. Jones, Kumaraswamy's distribution: A beta-type distribution with some tractability advantages, Stat. Methodol. 6 (2009) 70–81.

[31] S. Kotz, J. R. van Dorp, Beyond beta, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2004. Other continuous families of distributions with bounded support and applications.

[32] J. McCloskey, A Model for the Distribution of Individuals by Species in an Environment, Ph.D. thesis, Michigan State University, 1965.

[33] P. McCullagh, J. A. Nelder, Generalized linear models, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1989. Second edition [of MR0727836].

[34] R. B. Nelsen, An introduction to copulas, Springer Series in Statistics, Springer, New York, second edition, 2006.

[35] K. W. Ng, G.-L. Tian, M.-L. Tang, Dirichlet and related distributions, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester, 2011. Theory, methods and applications.

[36] V. Pawlowsky-Glahn, J. J. Egozcue, R. Tolosana-Delgado, Modeling and analysis of compositional data, Statistics in Practice, John Wiley & Sons, Ltd., Chichester, 2015.

[37] K. Pearson, 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 60 (1897) 489–498.

[38] R. Pyke, Spacings. (With discussion.), J. Roy. Statist. Soc. Ser. B 27 (1965) 395–449.

[39] J. S. Rao, M. Sobel, Incomplete Dirichlet integrals with applications to ordered uniform spacings, J. Multivariate Anal. 10 (1980) 603–610.

[40] A. Rényi, On the theory of order statistics, Acta Math. Acad. Sci. Hungar. 4 (1953) 191–231.

[41] M. Rosenblatt, Remarks on a multivariate transformation, Ann. Math. Statistics 23 (1952) 470–472.

[42] L. Rüschendorf, On the distributional transform, Sklar's theorem, and the empirical copula process, J. Statist. Plann. Inference 139 (2009) 3921–3927.

[43] T. Rychlik, Distributions and expectations of order statistics for possibly dependent random variables, J. Multivariate Anal. 48 (1994) 31–42.

[44] M. Sklar, Fonctions de répartition à *n* dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris 8 (1959) 229–231.

[45] A. Slavík, Product integration, its history and applications, volume 1 of Nečas Center for Mathematical Modeling, Matfyzpress, Prague, 2007. Corrected translation of the Czech original, Dějiny Matematiky/History of Mathematics, 29.

[46] P. V. Sukhatme, Tests of significance for samples of the x2-population with two degrees of freedom, Annals of Eugenics 8 (1937) 52–56.

[47] V. Volterra, Sui fondamenti della teoria delle equazioni differenziali lineari, Memorie della Società Italiana della Scienze 3 (1887).

[48] B. Wang, R. Wang, Joint mixability, Math. Oper. Res. 41 (2016) 808–826.

[49] R. Wang, Current open questions in complete mixability, Probability Surveys 12 (2015) 13 – 32.

[50] S. S. Wilks, Order statistics, Bull. Amer. Math. Soc. 54 (1948) 6–50.

[51] S. S. Wilks, Mathematical statistics, A Wiley Publication in Mathematical Statistics, John Wiley & Sons, Inc., New York-London, 1962.

[52] Wolfram Research, Inc., Mathematica, Version 14.0, 2024. Champaign, IL.