WORKING PAPERS

N° 1500

January 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 Representation of Compositional Data: Copulas and Regression models

Olivier P. Faugeras
Université Toulouse 1 Capitole - TSE

January 18, 2024

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 J. Aitchison (1986). We propose an alternative view of CoDa as stick breaking processes. 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. We establish connections with other topics of statistics like i) spacings and order statistics, ii) Bayesian nonparametrics and Dirichlet distributions, iii) neutrality, iv) mixability.

1 Introduction and summary

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, or vote shares in an election are examples of CoDa.

There are several competing ways to describe CoDa. The traditional approach (J. Aitchison (1986), Boogaart and Tolosana-Delgado (2013)) 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 $\mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d_{\geq 0}$ be a vector of non-negative absolute values of a composition, its closure is denoted as

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

This leads to the consideration of *normalised* (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[1]

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

CoDa points are then traditionally studied through a variety of log-ratio transforms, an approach pioneered by Aitchison John Aitchison (1982), J. Aitchison (1986). This gives rise to a special geometry, called Aitchison geometry, which turns the positive simplex into an Euclidean vector space. For recent accounts on the latter, see e.g., Greenacre (2018) or Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015). For an intrinsic approach to CoDa analysis based on projective geometry, see Faugeras (2023).

Inspired by Bayesian nonparametrics, we propose an alternative view of CoDa points, 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. 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).

The outline of the paper is as follows: in Section 2, we introduce our 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. Section 4 and 5 deal with statistical applications of such rescaled stick-breaking transformation for the study

[1] Note that we allow CoDa points with some null components, contrary to Aitchison's geometry(J. Aitchison (1986)), which is only defined on the interior of the simplex, so that log-coordinates ($\log p_i$), and the subsequent log-ratio transforms (alr,clr, ilr), are well-defined.

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 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, mixability, and CoDa with zeroes.

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

2.1 The ordered stick-breaking view

Our approaches are 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

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

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

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

(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.

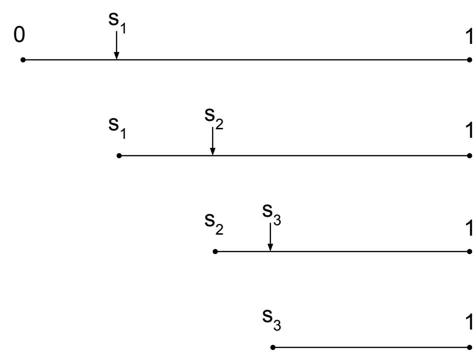

Figure 1: Stick-breaking of the unit interval

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

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

the corresponding order statistics. Eventually, define

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

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

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

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

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.

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

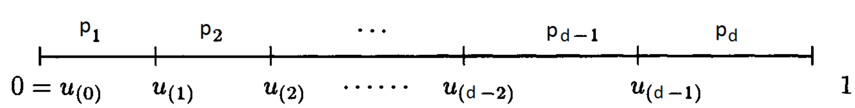

Figure 2: Stick-Ordered distribution $\mathtt{SO}$, obtained from the order statistics $u_{(i)}$

2.2.2 Examples

Example 1. *In particular, for $F = U_{[0,1]}$ the uniform distribution, (i.e. $F(x) = x$, $0 < x < 1$), $\mathbf{p} \sim \mathit{SO}(U_{[0,1]})$ gives a uniform distribution on the simplex, as* *shown in Pyke (1965) equation (2.1). (See also S. S. Wilks (1948) equation (6)). More precisely,* $(p_1, \dots, p_{d-1})$ *has a density w.r.t. the* $d-1$ *dimensional Lebesgue measure given by*

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

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*

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

which is symmetric in $(x_1,\dots,x_d)$ *(i.e. the* (p_i) *are exchangeable). One recognizes the Dirichlet* $\mathcal{Dir}(1,\dots,1;1)$ *distribution, see e.g. Rao and Sobel (1980) or Samuel S. Wilks (1962).*

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 Jones (2009)), or those of Kotz and Dorp (2004). (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 (Pyke (1965)).

Example 2 (Kumaraswamy)**.** *The Kumaraswamy distribution (Jones (2009))* $U \sim$ *Kumaraswamy*(α,β) *has density*

$$f(u) = \alpha\beta u^{\alpha-1}(1-u^{\alpha})^{\beta-1}, \quad 0 < u < 1, \tag{4}$$

and cdf

$$F(u) = 1 - (1-x^{\alpha})^{\beta}, \quad 1 < u < 1. \tag{5}$$

Moreover, for i.i.d. $u_i \sim F$*,* $i = 1,\dots,d-1$*, with density* f*, the marginal distribution of the spacings writes (see e.g. Pyke (1965) p. 399)*

$$f_{p_i}(x) = \frac{(d-1)!}{(i-2)!(d-1-i)!} \int (F(t))^{i-2}(1-F(x+t))^{d-1-i} f(t) f(x+t) dt, \tag{6}$$

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

$$f_{p_1}(x) = f_{u_{(1)}}(x) = (d-1) f(x)(1-F(x))^{d-2}. \tag{7}$$

Applying formulas (6) and (7) to (4) and (5) 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 (2.1) by taking dependent r.v. u_i instead of independent ones.

Definition 2.2 (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* $\mathbf{p} \sim \mathtt{GSO}(F_{\mathbf{u}})$*, viz.*

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

For non-identically distributed or dependent variables, one can compute them e.g. using results of David and Nagaraja (2003) Chap. 5. Jaworski and Rychlik (2008), Rychlik (1994) (See also Balakrishnan N. (1998)). 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* $p_1 = p_3$*.*

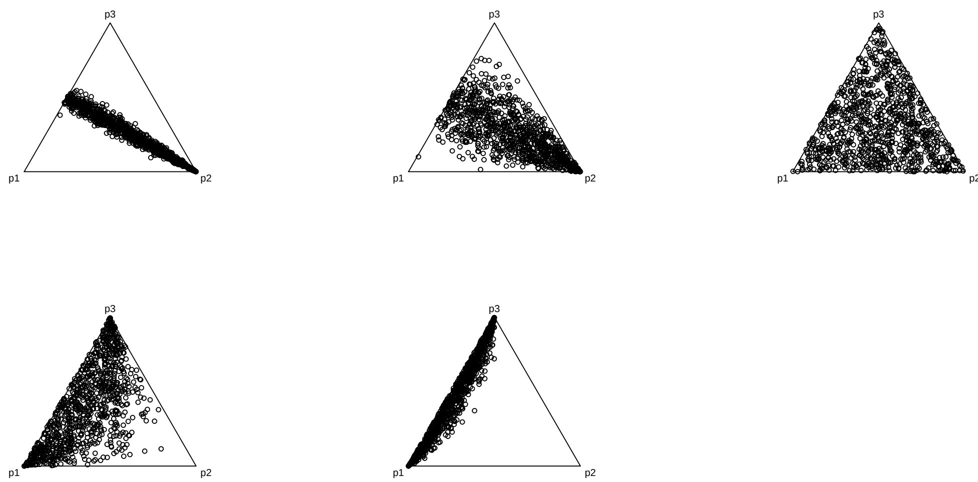

Figure 3: Ternary plots of Generalized-Stick-Ordered distribution for $d = 3$, with Gaussian copula generator, with varying correlation coefficient ρ. From left to right and up to down: $\rho = -0.99$, $\rho = -0.8$, $\rho = 0$, $\rho = 0.8$, $\rho = 0.99$.

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 propose 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 Halmos (1944), Connor and Mosimann (1969) among others, it is therefore natural to introduce the transformation,

$$
\begin{aligned}
z_1 &= s_1 = p_1, \\
z_i &= p_i/(1 - s_{i-1}), \quad i = 1, \ldots, d-1, \\
z_d &= 1,
\end{aligned}
\tag{8}
$$

with inverse transformation,

$$p_1 = z_1,$$
$$p_i = z_i \prod_{j=1}^{i-1}(1 - z_j), \quad i = 2, \ldots, d-1,$$
$$p_d = \prod_{i=1}^{d-1}(1 - z_i). \tag{9}$$

By construction, the z_i in (8) 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 (8) turns the "akward" simplex Δ^{d-1} into the unit cube $[0,1]^{d-1}$ (we drop $z_d = 1$ as it is equal to 1). This leads to a free unit cube view of CoDa points.

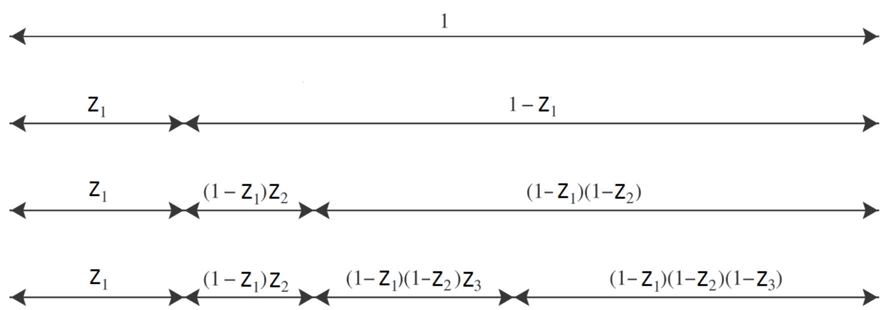

Figure 4: Rescaled stick-breaking

Remark 2. *Historically, the transformation (8) was introduced by Halmos (1944) 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 (8) from a sequence of independent* $(z_i) \in [0,1]$ *is called a residual allocation model (RAM), see Feng (2021) 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 Griffiths (1979), Engen (1975), McCloskey (1965), see e.g. Ewens (1990). 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 Samuel S. Wilks (1962) p. 178, Ferguson (1973), Ishwaran and James (2001). 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 Connor and

Mosimann (1969), 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

$$\left(\frac{p_2}{1-p_1}, \quad \ldots, \quad \frac{p_d}{1-p_1} \right), \tag{10}$$

and if p_1 is independent of the latter rescaled subcomposition, one can eliminate p_1 from the analysis of $\mathbf{p}$. Therefore, Connor and Mosimann (1969) defines neutrality as follows: p_1 is said to be neutral if p_1 is independent of (10): 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

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

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 Connor and Mosimann (1969)). A further generalization of neutrality is complete neutrality: if the $\mathbf{z} = (z_1, \ldots, z_{d-1})$ of (8) are mutually independent, then the corresponding $\mathbf{p}$ is said to be completely neutral, or equivalently (Theorem 2 in Connor and Mosimann (1969)) 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 (8) to obtain a completely neutral distribution on the simplex. In particular, Connor and Mosimann (1969) construct a generalisation of the Dirichlet distribution from independent $z_i \sim \beta(a_i, b_i)$. Ng, Tian, and Tang (2011) Theorem 2.2 use the transformations (8) and (9) 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 Ferguson (1973), see e.g. Ghosal and Vaart (2017) 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 (9).

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 (9), i.e. s.t.

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

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 $(Z_1, \ldots, Z_d)$ are independent with $Z_i \sim \beta(\alpha_i, \sum_{j>i} \alpha_j)$, then $\mathbf{p}$ is Dirichlet $\texttt{Dir}(k; \alpha_1, \ldots, \alpha_d)$ distributed, see Ghosal and Vaart (2017) Corollary G.5.

3.4 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 $\mathbf{z}$ 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 $\mathbf{p}$.

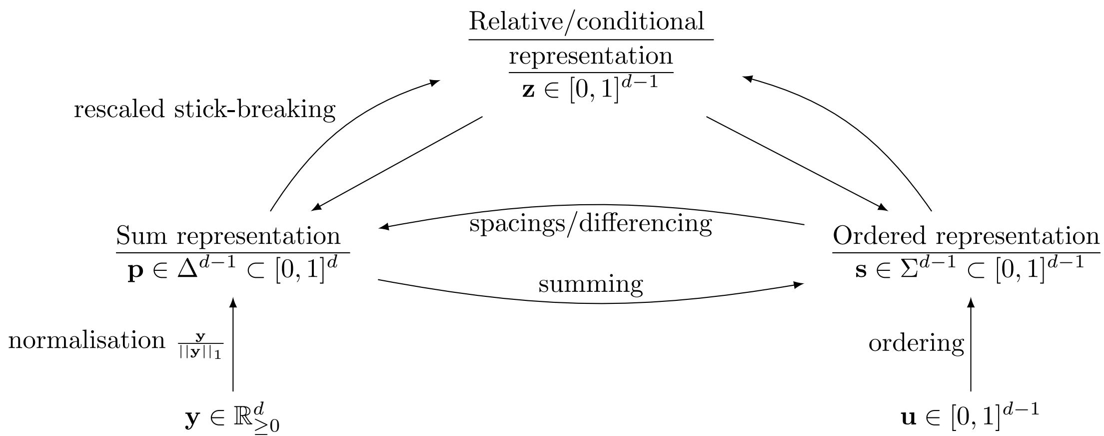

Figure 5: Representations of the simplex

In addition, Figure 5 shows how one can obtain the stick-ordered distribution of Definition 2.1, 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 $\mathcal{C}$, i.e. normalisation by the sum of nonnegative random variables, for some $\mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d_{\geq 0}$. 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 Sukhatme (1937), Rényi (1953), Pyke (1965)) 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 $s_i = u_{(i)}$,

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

and

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

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 Samuel S. Wilks (1962) p. 178, 182 and 238.

3.5 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 (8). 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 $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{d-1}) \in \mathbb{R}^{d-1}$ by applying an increasing[2] 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).

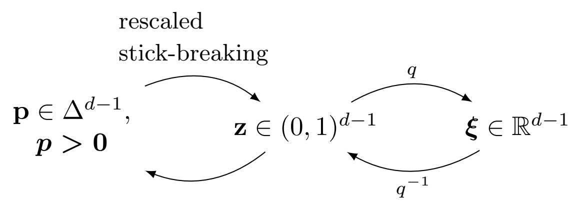

Figure 6: Free Euclidean representation $\boldsymbol{\xi} \in \mathbb{R}^{d-1}$ of positive CoDa.

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. 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 $\mathbf{p} \in \Delta^{d-1}$. Potential applications are numerous.

4 Application of unit cube geometry: Copulas for CoDa

In the complete neutrality view of Connor and Mosimann (1969), 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 (Nelsen (2006)) 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.

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

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

By back transformation (9), 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 $Q_{z_i} = F_{z_i}^{-1} : [0,1] \to [0,1]$, $i = 1, \ldots, d-1$, 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, (9) yields

$$p_1 = Q_{z_1}(v_1), \tag{11}$$

$$p_i = Q_{z_i}(v_i) \prod_{j=1}^{i-1} (1 - Q_{z_j}(v_j)), \quad i = 2, \ldots, d-1, \tag{12}$$

$$p_d = \prod_{i=1}^{d-1} (1 - Q_{z_i}(v_i)). \tag{13}$$

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 (8), 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.

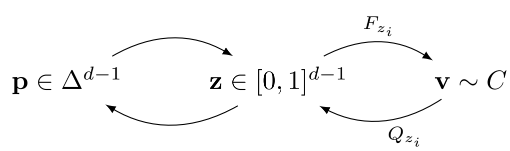

Figure 7: Probabilistic construction of a CoDapula

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

$$F_{z_i}(x, \eta) := P(z_i < x) + \eta P(z_i = x), \quad \eta \in [0,1]. \tag{14}$$

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. Ludger Rüschendorf (2009), Faugeras and Ludger Rüschendorf (2017)).

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

Definition 4.1 (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 (8). 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 (Sklar (1959)), a CoDapula always exists. It is unique if $\mathbf{z}$ is continuous (see e.g. Nelsen (2006)). Definition 4.1 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 (11), 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 4.1 and marginal quantile functions for $\mathbf{z}$. In Figure 8, the copula of $\mathbf{z}$ 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 $\mathbf{z}$ domains, and the right panels the corresponding ternary plots in the simplex CoDa space for $\mathbf{p}$. 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 $\mathbf{z}$ level, resulting in the pattern of points shown on the right panel at the CoDa level. More examples could be considered.

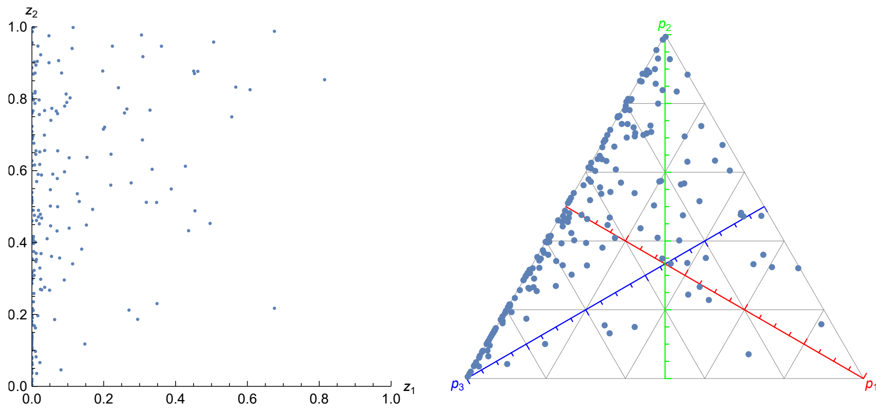

Figure 8: CoDa with AMH CoDapula ($\alpha = 0.91$), and $\beta(1/4, 2)$, $U_{[0,1]}$ marginal distribution functions: scatter plot at the $\mathbf{z}$ level (left) and ternary scatter plot for the resulting $\mathbf{p}$ (right), $d = 3$.

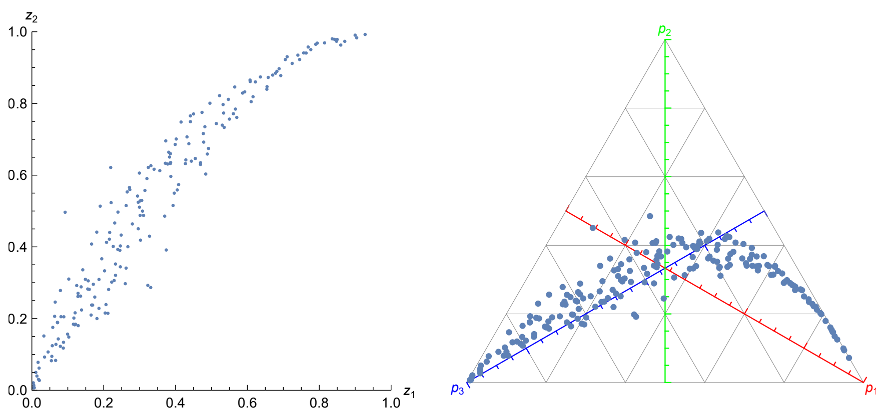

Figure 9: CoDa with Gumbel CoDapula ($\rho = 7$) and $\beta(1/4, 2)$, $U_{[0,1]}$ marginal distribution functions; scatter plot at the $\mathbf{z}$ level (left) and ternary scatter plot for the resulting $\mathbf{p}$ (right), $d = 3$.

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 $\mathbf{z}$ means that $\mathbf{p}$ is completely neutral.

At another extreme, complete dependence at the z level induces a specific dependence pattern at the $\mathbf{p}$ 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 $M(x_1, \ldots, x_{d-1}) = \min(x_1, \ldots, x_{d-1})$. The comonotone copula corresponds to the distribution of the vector $\mathbf{v} = (v, \ldots, v) \in [0,1]^{d-1}$, with a single $v \sim U_{[0,1}$. In other words,

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

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

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

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

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

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

$$\begin{aligned} p_1 &= Q_{z_1}(v) \\ p_2 &= Q_{z_2}(v)(1 - Q_{z_1}(v)) \\ p_3 &= 1 - p_1 - p_2 = (1 - Q_{z_1}(v))(1 - Q_{z_2}(v)) \end{aligned} \tag{15}$$

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 (15), so that $p_1 = v$, $p_2 = v(1 - v)$, $p_3 = (1 - v)^2$.

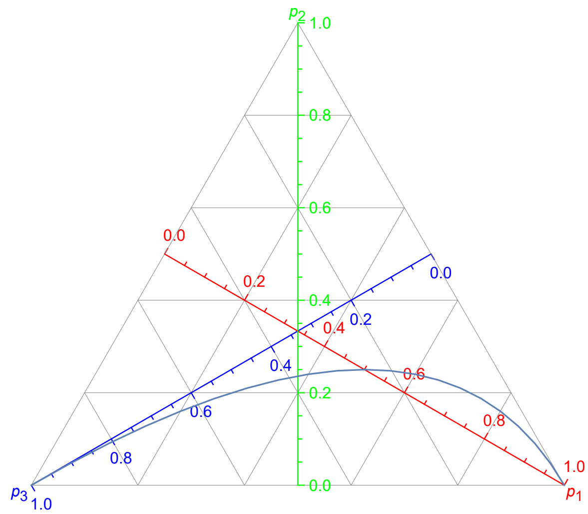

Figure 10: Ternary plot of a CoDa with comonotone CoDapula and uniform quantile functions, $d = 3$, with barycentric axes p_1 (red), p_2 (green), p_3 (blue).

The distribution of $\mathbf{p}$ *is singular, as each component* p_i *is a deterministic function of* $v \sim U_{[0,1]}$*:* $\mathbf{p}$ *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* $p_2 \in [0, 1/4]$.

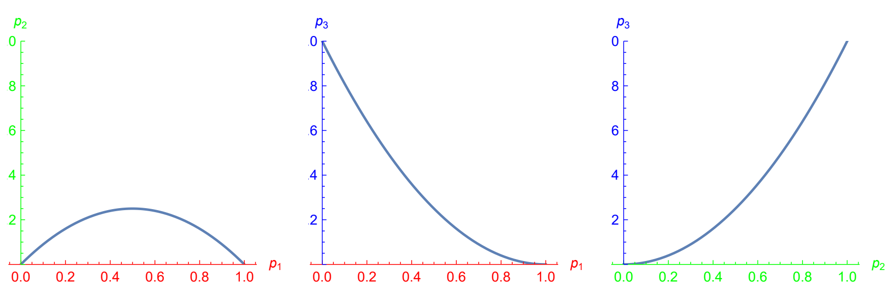

Figure 11: Complete dependence between pairs of components of a CoDa with comonotone CoDapula and uniform quantile functions, $d = 3$: (p_1, p_2) (left), counter-monotone (p_1, p_3) (middle), comonotone (p_2, p_3) (right)

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

$$\mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} Q_{z_1}(v) \\ Q_{z_2}(1-v)(1-Q_{z_1}(v)) \\ (1-Q_{z_1}(v))(1-Q_{z_2}(1-v)) \end{pmatrix}. \tag{16}$$

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

Example 5 (Counter-monotone CoDapula)**.** *For $Q_{z_1}(v) = Q_{z_2}(v) = v$, (16) 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.*

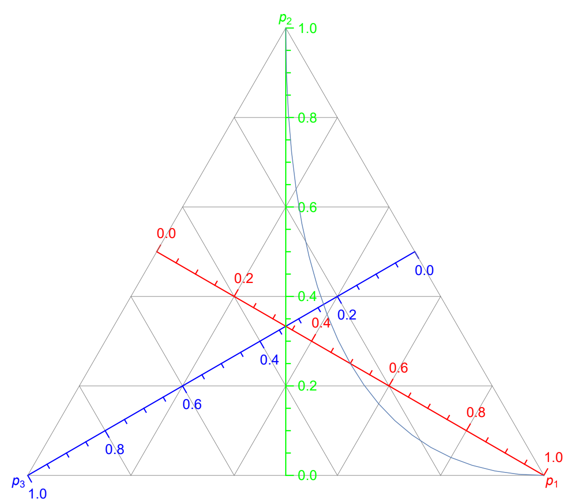

Figure 12: Ternary plot of a CoDa with countermonotone CoDapula and uniform quantile functions, $d = 3$, with barycentric axes p_1 (red), p_2 (green), p_3 (blue).

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).*

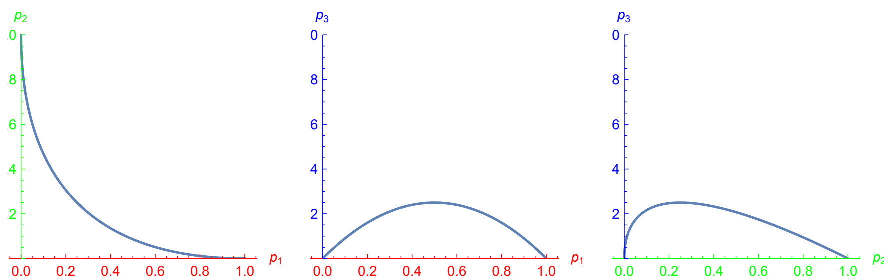

Figure 13: Complete dependence between pairs of components of a CoDa with countermonotone CoDapula and uniform quantile functions, $d = 3$: countermonotone (p_1, p_2) (left), (p_1, p_3) (middle), (p_2, p_3) (right)

5 Application of unit cube geometry: Regression models for CoDa

As a second possible generalization, the rescaled stick-breaking approach (8) 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 (8) 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 $\mathbf{X} = (X_1, \ldots X_k) \in \mathbb{R}^k$ be a vector with joint c.d.f. F. If one can transform $\mathbf{X}$ into a sequence $\epsilon_1, \ldots, \epsilon_k$ 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

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

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 $\mathbf{X}$.

Rosenblatt (1952)'s transform and its generalizations (see Ludger Rüschendorf (2009)) achieves such a reduction: Denote by $F_{i|i-1,\ldots,1}$ the conditional c.d.f. of X_i

given $(X_{i-1}, \ldots, X_1)$, $i = 2, \ldots, k$, with F_1 the (marginal) cdf of X_1. Rosenblatt (1952)'s transform is then defined by $\boldsymbol{\epsilon} = (\epsilon_1, \ldots \epsilon_k)$ with

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

Under an assumption of continuity[3] of the successive conditional c.d.f. $F_{i|i-1,\ldots,1}$, Rosenblatt's transform turn the vector $\mathbf{X}$ into a vector $\boldsymbol{\epsilon}$ of i.i.d. $U_{[0,1]}$ components (see Ludger Rüschendorf (2009)).

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

$$\begin{aligned} X_1 &:= F_1^{-1}(\epsilon_1) \\ X_i &:= F^{-1}_{i|i-1,\ldots,1}(\epsilon_i | X_{i-1}, \ldots, X_1), \quad i = 2, \ldots, k, \end{aligned} \tag{17}$$

one obtains a vector $\mathbf{X}$ with the desired joint c.d.f. F. Each equation (17) 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 Ludger Rüschendorf (2009), where (17) is a (triangular) stochastic representation of the successive predictive distributions $P^{X_i|X_{i-1},\ldots,X_1}$.

In (17), 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 Faugeras and Ludger Rüschendorf (2017)). 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 (17) 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 (17) to the z_i of (8) 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

[3] For discontinuous conditional cdf, one must use the conditional probability integral transform (14) instead, see Ludger Rüschendorf (2009)

$i = 1, \ldots, d-1$. Then, by back transformation (8), 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

$$\begin{aligned} z_1 &= \phi_1(\epsilon_1) \\ z_i &= \phi_i(\epsilon_i, z_{i-1}, \ldots, z_1), \quad i = 2, \ldots, d-1, \end{aligned} \tag{18}$$

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

$$\begin{aligned} z_1 &= \Phi(\epsilon_1) \\ z_i &= \Phi(a_{i,1} z_1 + \ldots + a_{i,i-1} z_{i-1} + \epsilon_i), \quad i = 2, \ldots, d-1 \end{aligned} \tag{19}$$

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. McCullagh and Nelder (1989), and more specifically, Bonat, Ribeiro Jr, and Zeviani (2012) 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 (Wolfram Research, Inc. (n.d.)) "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 Wolfram Research, Inc. (n.d.)) 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,

$$p_1 = p_1$$
$$p_2 \approx (1 - p_1)(0.5 + 0.047 \cos(-21.86\, p_1 + 8.76))$$
$$p_3 \approx 1 - p_1 - p_2 \approx (1 - p_1)(0.5 - 0.047 \cos(-21.86\, p_1 + 8.76)).$$

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

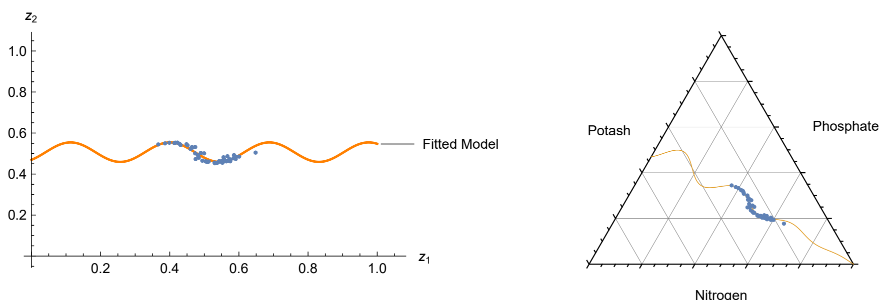

Figure 14: Internal parametric regression model for agriculture data. Scatter plots (blue points) and fitted sinusoidal model (orange line) in the **z** space (left) and for the original data (right).

5.4 Extensions and alternatives

As shown in the previous example of Section 5.3, the rescaled stick-breaking transformation (8) 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 **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 (17) 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 (8) 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 3.5. 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 (8), (i.e. the functions ϕ_i in (18) or the coefficients $a_{j,i}$ in (19)) 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 (9).

- 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 Griessenberger, Junker, and Trutschnig (2022)'s $\zeta^1(\mathbf{X}, Y)$ or Azadkia and Chatterjee (2021)'s $T(\mathbf{X}, Y)$[4], the latter being a multivariate extension of the bivariate measure ξ of Chatterjee (2020). These coefficients quantifies the extent of regression dependence of a univariate random variable Y on a k-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_k)$: 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 $\zeta^1((z_2, \ldots, z_d), z_1)$ or $T((z_2, \ldots, z_d), z_1)$.

6 Conclusion and further remarks

6.1 Conclusion

We have proposed 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 Halmos (1944) and Connor and Mosimann (1969), removes the unit-sum constraint of the simplex representation and turns a CoDa point into a free vector of the unit cube. 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. Such an approach is an interesting

[4]Note that Azadkia and Chatterjee (2021) introduce a more general regression dependence coefficient which allows for covariates and the assessment of *conditional* independence.

alternative to the classical log-ratio coordonatizations techniques of Aitchison and his followers. In particular, we introduced the concept of a CoDapula, a copula for CoDa, and showed how to study the intra-dependence of CoDa with such copulas or via regression techniques.

For length reasons, we have barely scratched the surface of statistical applications based on these transformations. Much more needs to be done to explore the potentialities of the proposed approaches. Being now reduced to vector data, 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. Let us close the article with some further remarks.

6.2 Choice of the ordering of the components

The only possible issue of the transformations (2) and (8) 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.

- 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 (9), between the first component p_1 and the remaining ones $p_2, \dots, p_d$: 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 $p_2, \dots, p_d$ 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 $z_1 = r(z_2, \dots, z_d, \epsilon)$ 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. $p_1 = r(z_2, \dots, z_d, \epsilon)$.

- 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 Griessenberger, Junker, and Trutschnig (2022) or $T(\mathbf{X}, Y)$ of Azadkia and Chatterjee (2021), 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 proposed approaches were constructive and gave explicit representations of distributions of CoDa points. In case of the ordered approach, the stick-ordered distributions of Definition 2.1, 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 R. Wang (2015)). This connection between mixability and distributions for CoDa does not seem to have been made beforehand by the CoDa community.

The definition of joint mixability (B. Wang and R. Wang (2016)) is as follows:

Definition 6.1. *An d-tuple of probability distributions on $\mathbb{R}$, $(F_1, \dots, F_d)$ is jointly mixable if there exist d random variables $X_1 \sim F_1, \dots, X_d \sim F_d$ such that $X_1 + \dots + X_d =: K$ 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 $K = 1$. Gaffke and L. Rüschendorf (1981) Theorem 5 give a necessary and sufficient condition. Necessary conditions are given in Theorem 2.1 in B. Wang and R. Wang (2016), and sufficient conditions are given in Theorems 3.1, 3.2, and 3.4 for uniform, monotone, and symmetric-unimodal densities, respectively.

6.4 CoDa with some zero components

The well-known additive log-ratio transformation (alr) of J. Aitchison (1986)

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

and its variants clr and ilr (Greenacre (2018), Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015)) turns a CoDa point $\mathbf{p}$ as a vector element of $\mathbb{R}^{d-1}$. However, it is undefined if $\mathbf{p}$ has some zero components and special treatments of the zero components are required, like amalgamation of the finer parts, replacement of the zeroes with small values, or treatment of the zero observations as outliers. However, these treatments have an ad-hoc character and are not completely satisfactory from a modeling perspective. Hence, the statistical literature based on log-ratio analysis usually enforces an assumption of strict non-negativity of the components, which limits its scope of application.

To the contrary, the stick-breaking representations do not require such an assumption: 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 2.1 of the ordered representation (3) have common discrete components in their distributions. For the rescaled $\mathbf{z}$ representation (8), 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.

Acknowledgements

[author] 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

Aitchison, J. (1986). *The statistical analysis of compositional data.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, pp. xvi+416. ISBN: 0-412-28060-4. DOI: 10.1007/978-94-009-4109-0. URL: https://doi.org/10.1007/978-94-009-4109-0.

Aitchison, John (1982). "The statistical analysis of compositional data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44(2), pp. 139–160.

Azadkia, Mona and Sourav Chatterjee (2021). "A simple measure of conditional dependence". In: *Ann. Statist.* 49(6), pp. 3070–3102. ISSN: 0090-5364. DOI: 10.1214/21-aos2073. URL: https://doi.org/10.1214/21-aos2073.

Balakrishnan N., Rao C.R. (eds.) (1998). *Handbook of statistics 16. Order statistics: theory and methods.* Vol. 16. Elsevier Science. ISBN: 0444820914,9780444820914.

Bonat, Wagner Hugo, PJ Ribeiro Jr, and Walmes Marques Zeviani (2012). "Regression models with responses on the unity Interval: Specification, estimation and comparison". In: *Biometric Brazilian Journal* 30(4), pp. 415–431.

Boogaart, K. Gerald van den and Raimon Tolosana-Delgado (2013). *Analyzing compositional data with R.* Use R! Springer, Heidelberg, pp. xvi+258. ISBN: 978-3-642-36808-0; 978-3-642-36809-7. DOI: 10.1007/978-3-642-36809-7. URL: https://doi.org/10.1007/978-3-642-36809-7.

Chatterjee, Sourav (2020). "A New Coefficient of Correlation". In: *Journal of the American Statistical Association* 0(0), pp. 1–21. DOI: 10.1080/01621459.2020.1758115. eprint: https://doi.org/10.1080/01621459.2020.1758115. URL: https://doi.org/10.1080/01621459.2020.1758115.

Connor, Robert J. and James E. Mosimann (1969). "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution". In: *Journal of the American Statistical Association* 64(325), pp. 194–206. DOI: 10.1080/01621459.1969.10500963. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1969.10500963. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10500963.

David, H. A. and H. N. Nagaraja (2003). *Order statistics.* Third. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, pp. xvi+458. ISBN: 0-471-38926-9. DOI: 10.1002/0471722162. URL: https://doi.org/10.1002/0471722162.

Engen, Steiner (1975). "A note on the geometric series as a species frequency model". In: *Biometrika* 62(3), pp. 697–699. ISSN: 0006-3444. DOI: 10.1093/biomet/62.3.697. URL: https://doi.org/10.1093/biomet/62.3.697.

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

Faugeras, Olivier P. (2023). "An invitation to intrinsic compositional data analysis using projective geometry and Hilbert's metric". In: *TSE Working Paper, no. 23-1496.* URL: https://www.tse-fr.eu/fr/publications/

invitation-intrinsic-compositional-data-analysis-using-projective-geometry-and-hilberts-metric.

Faugeras, Olivier P. and Ludger Rüschendorf (2017). "Markov morphisms: a combined copula and mass transportation approach to multivariate quantiles". In: *Math. Appl. (Warsaw)* 45(1), pp. 21–63. ISSN: 1730-2668. DOI: 10.14708/ma.v45i1.2921. URL: https://doi.org/10.14708/ma.v45i1.2921.

Feng, Shui (2021). "A note on residual allocation models". In: *Front. Math. China* 16(2), pp. 381–394. ISSN: 1673-3452. DOI: 10.1007/s11464-020-0871-8. URL: https://doi.org/10.1007/s11464-020-0871-8.

Ferguson, Thomas S. (1973). "A Bayesian analysis of some nonparametric problems". In: *Ann. Statist.* 1, pp. 209–230. ISSN: 0090-5364. URL: http://links.jstor.org/sici?sici=0090-5364(197303)1:2%3C209:ABAOSN%3E2.0.CO;2-U&origin=MSN.

Gaffke, N. and L. Rüschendorf (1981). "On a class of extremal problems in statistics". In: *Math. Operationsforsch. Statist. Ser. Optim.* 12(1), pp. 123–135. ISSN: 0323-3898. DOI: 10.1080/02331938108842712. URL: https://doi.org/10.1080/02331938108842712.

Ghosal, Subhashis and Aad van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, pp. xxiv+646. ISBN: 978-0-521-87826-5. DOI: 10.1017/9781139029834. URL: https://doi.org/10.1017/9781139029834.

Greenacre, Michael (2018). *Compositional data analysis in practice*. Chapman and Hall/CRC.

Griessenberger, Florian, Robert R. Junker, and Wolfgang Trutschnig (2022). "On a multivariate copula-based dependence measure and its estimation". In: *Electron. J. Stat.* 16(1), pp. 2206–2251. DOI: 10.1214/22-ejs2005. URL: https://doi.org/10.1214/22-ejs2005.

Griffiths, R. C. (1979). "Exact sampling distributions from the infinite neutral alleles model". In: *Adv. in Appl. Probab.* 11(2), pp. 326–354. ISSN: 0001-8678. DOI: 10.2307/1426843. URL: https://doi.org/10.2307/1426843.

Halmos, Paul R. (1944). "Random alms". In: *Ann. Math. Statistics* 15, pp. 182–189. ISSN: 0003-4851. DOI: 10.1214/aoms/1177731283. URL: https://doi.org/10.1214/aoms/1177731283.

Ishwaran, Hemant and Lancelot F. James (2001). "Gibbs sampling methods for stick-breaking priors". In: *J. Amer. Statist. Assoc.* 96(453), pp. 161–173. ISSN: 0162-1459. DOI: 10.1198/016214501750332758. URL: https://doi.org/10.1198/016214501750332758.

Jaworski, Piotr and Tomasz Rychlik (2008). "On distributions of order statistics for absolutely continuous copulas with applications to reliability". In: *Kybernetika (Prague)* 44(6), pp. 757–776. ISSN: 0023-5954.

Jones, M. C. (2009). "Kumaraswamy's distribution: A beta-type distribution with some tractability advantages". In: *Stat. Methodol.* 6(1), pp. 70–81. ISSN: 1572-3127. DOI: 10.1016/j.stamet.2008.04.001. URL: https://doi.org/10.1016/j.stamet.2008.04.001.

Kotz, Samuel and Johan René van Dorp (2004). *Beyond beta*. Other continuous families of distributions with bounded support and applications. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, pp. xvi+289. ISBN: 981-256-115-3. DOI: 10.1142/5720. URL: https://doi.org/10.1142/5720.

McCloskey, JW (1965). "A Model for the Distribution of Individuals by Species in an Environment, unpublished Ph. D". PhD thesis. PhD thesis, Michigan State University.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Second edition [of MR0727836]. Chapman & Hall, London, pp. xix+511. ISBN: 0-412-31760-5. DOI: 10.1007/978-1-4899-3242-6. URL: https://doi.org/10.1007/978-1-4899-3242-6.

Nelsen, Roger B. (2006). *An introduction to copulas*. Second. Springer Series in Statistics. Springer, New York, pp. xiv+269. ISBN: 978-0387-28659-4; 0-387-28659-4. DOI: 10.1007/s11229-005-3715-x. URL: https://doi.org/10.1007/s11229-005-3715-x.

Ng, Kai Wang, Guo-Liang Tian, and Man-Lai Tang (2011). *Dirichlet and related distributions*. Wiley Series in Probability and Statistics. Theory, methods and applications. John Wiley & Sons, Ltd., Chichester, pp. xxvi+310. ISBN: 978-0-470-68819-9. DOI: 10.1002/9781119995784. URL: https://doi.org/10.1002/9781119995784.

Pawlowsky-Glahn, Vera, Juan José Egozcue, and Raimon Tolosana-Delgado (2015). *Modeling and analysis of compositional data*. Statistics in Practice. John Wiley & Sons, Ltd., Chichester, pp. xx+247. ISBN: 978-1-118-44306-4.

Pyke, R. (1965). "Spacings. (With discussion.)" In: *J. Roy. Statist. Soc. Ser. B* 27, pp. 395–449. ISSN: 0035-9246. URL: http://links.jstor.org/sici?sici=0035-9246(1965)27:3%3C395:S%3E2.0.CO;2-C&origin=MSN.

Rao, J. S. and Milton Sobel (1980). "Incomplete Dirichlet integrals with applications to ordered uniform spacings". In: *J. Multivariate Anal.* 10(4), pp. 603–610. ISSN: 0047-259X. DOI: 10.1016/0047-259X(80)90073-1. URL: https://doi.org/10.1016/0047-259X(80)90073-1.

Rényi, Alfréd (1953). "On the theory of order statistics". In: *Acta Math. Acad. Sci. Hungar.* 4, pp. 191–231. ISSN: 0001-5954. DOI: 10.1007/BF02127580. URL: https://doi.org/10.1007/BF02127580.

Rosenblatt, Murray (1952). "Remarks on a multivariate transformation". In: *Ann. Math. Statistics* 23, pp. 470–472. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729394. URL: https://doi.org/10.1214/aoms/1177729394.

Rüschendorf, Ludger (2009). "On the distributional transform, Sklar's theorem, and the empirical copula process". In: *J. Statist. Plann. Inference* 139(11), pp. 3921–3927. ISSN: 0378-3758. DOI: 10.1016/j.jspi.2009.05.030. URL: https://doi.org/10.1016/j.jspi.2009.05.030.

Rychlik, Tomasz (1994). "Distributions and expectations of order statistics for possibly dependent random variables". In: *J. Multivariate Anal.* 48(1), pp. 31–42. ISSN: 0047-259X. DOI: 10.1016/0047-259X(94)80003-E. URL: https://doi.org/10.1016/0047-259X(94)80003-E.

Sklar, M. (1959). "Fonctions de répartition à n dimensions et leurs marges". In: *Publ. Inst. Statist. Univ. Paris* 8, pp. 229–231.

Sukhatme, Pandurang V (1937). "TESTS OF SIGNIFICANCE FOR SAMPLES OF THE x2-POPULATION WITH TWO DEGREES OF FREEDOM". In: *Annals of Eugenics* 8(1), pp. 52–56.

Wang, Bin and Ruodu Wang (2016). "Joint mixability". In: *Math. Oper. Res.* 41(3), pp. 808–826. ISSN: 0364-765X. DOI: 10.1287/moor.2015.0755. URL: https://doi.org/10.1287/moor.2015.0755.

Wang, Ruodu (2015). "Current open questions in complete mixability". In: *Probability Surveys* 12(none), pp. 13–32. DOI: 10.1214/14-PS250. URL: https://doi.org/10.1214/14-PS250.

Wilks, S. S. (1948). "Order statistics". In: *Bull. Amer. Math. Soc.* 54, pp. 6–50. ISSN: 0002-9904. DOI: 10.1090/S0002-9904-1948-08936-4. URL: https://doi.org/10.1090/S0002-9904-1948-08936-4.

Wilks, Samuel S. (1962). *Mathematical statistics.* A Wiley Publication in Mathematical Statistics. John Wiley & Sons, Inc., New York-London, pp. xvi+644.

Wolfram Research, Inc. (n.d.). *Mathematica, Version 14.0.* Champaign, IL, 2024.