WORKING PAPERS

N° 1744

May 2026

“Coherent Ratios for Compositional Data Analysis with Zeros”

Olivier Faugeras

Toulouse School of Economics

COHERENT RATIOS FOR COMPOSITIONAL DATA ANALYSIS WITH ZEROS

OLIVIER P. FAUGERAS

Toulouse School of Economics, Université de Toulouse Capitole
Institut de Mathématiques de Toulouse, France

Abstract. Compositional data (CoDa) are scale-invariant by nature, and their analysis traditionally relies on log-ratio transformations. A cornerstone of the Aitchison school is the principle of subcompositional coherence: analyses should be consistent when focusing on any subset of parts, a property that follows from the isomorphism property of the logarithm. However, zeros in the data render log-ratio transformations undefined, posing a fundamental challenge.

This paper introduces a family of transformations based on a ratio-based homeomorphism, which maps extended non-negative numbers $[0, \infty]$ to the unit interval. By applying it to suitable ratios of components, we construct both i) *local* representations that require a reference component (or a set of components) to be non-zero, and ii) *global* representations that handle zeros in all components simultaneously. We show that the local representations preserve subcompositional coherence for subcompositions containing the reference part, a property deemed essential by the log-ratio community, while the global representations trade this coherence for the ability to accommodate zeros everywhere. Thus, no single transformation satisfies all desirable properties, but the local approach offers a principled compromise: it retains subcompositional coherence and enables a quasi-global treatment of zeros.

The practical utility of the approach is illustrated on a glass dataset for predicting the refractive index, where the proposed local representation outperforms standard log-ratio methods with zero imputation, matches industry benchmarks, and yields interpretable coefficients consistent with domain knowledge. The proposed framework provides a flexible, imputation-free toolbox for analyzing compositional data with zeros, allowing analysts to choose between coherence and full coverage depending on the application, and enabling the use of standard multivariate techniques on bounded, interpretable coordinates.

1. Introduction

1.1. **Motivation and notation.** Compositional data arise in a wide range of scientific disciplines, including geochemistry, ecology, microbiomics, and economics. They consist of multivariate observations representing the relative contributions of different components to a whole. Formally, set $\mathbb{R}_+ := \{x \in \mathbb{R}, x \geq 0\}$,

E-mail address: olivier.faugeras@tse-fr.eu.

Date: April 8, 2026.

2020 *Mathematics Subject Classification.* 62R20, 62H20.

Key words and phrases. compositional data, zeroes, subcompositional coherence.

resp. $\mathbb{R}_{++} := \{x \in \mathbb{R}, x > 0\}$ the non-negative, resp. positive reals, and let $\mathbf{u} = (u_0, u_1, \ldots, u_d) \in \mathbb{R}_+^{d+1} \setminus \{\mathbf{0}\}$ be a vector of non-negative raw measurements (e.g. counts). Its projective/compositional part (Faugeras, 2023) corresponds to the direction of $\mathbf{u}$, i.e. equivalence classes $[\mathbf{u}]_+ := \{\lambda\mathbf{u}, \lambda > 0\}$ for the scaling relation. The corresponding set of equivalence classes

$$\mathbb{P}_+^d := \{[\mathbf{u}]_+, \mathbf{u} \in \mathbb{R}_+^{d+1} \setminus \{\mathbf{0}\}\}$$

is defined as the CoDa space (thus possibly with zeroes in the components of $\mathbf{u}$).

More classically, such (projective) CoDa space is often represented via its affine model, that is the simplex

$$\Delta_+^d := \left\{\mathbf{x} \in \mathbb{R}_+^d \setminus \{\mathbf{0}\} : \sum_{i=0}^{d} x_i = 1\right\},$$

obtained by ℓ_1 normalization (called closure in the CoDa literature) of any representative $\mathbf{u}$ of $[\mathbf{u}]_+$: set

$$\mathbf{x} := \frac{\mathbf{u}}{\|\mathbf{u}\|_1} =: \mathcal{C}(\mathbf{u}) \in \Delta_+^d \tag{1}$$

the simplex representative of $[\mathbf{u}]_+$, where $\|\mathbf{u}\|_1 = \sum_{i=0}^{d} |u_i|$ is the ℓ_1 norm.

Hence, a $d+1$-part composition is often thought directly as a vector in the d-dimensional simplex, which is a constrained subset of Euclidean space. Because of the unit-sum constraint, standard multivariate statistical methods that assume an unconstrained Euclidean sample space are generally inappropriate for compositional data. Applying classical methods directly can lead to artifacts such as spurious correlations (Pearson, 1897) and misleading inferential conclusions. To address these issues, Aitchison, 1982, 1986 developed a rigorous statistical framework for compositional data analysis based on log-ratio transformations.

The fundamental log-ratio transformations—namely the *additive log-ratio* (alr), *centered log-ratio* (clr), and *isometric log-ratio* (ilr)—rely on the logarithms of component ratios to yield isomorphic representations of the open simplex. They effectively map the *positive*[1] simplex, defined as $\Delta_{++}^d := \left\{\mathbf{x} \in \mathbb{R}_{++}^d : \sum_{i=0}^{d} x_i = 1\right\}$, into an unconstrained Euclidean space. Consequently, the positive simplex is endowed with a formal Hilbert space structure, $(\Delta_{++}^d, \oplus, \odot, \langle\cdot,\cdot\rangle_A)$. In this geometry, $\oplus$ and $\odot$ represent the vector space operations of perturbation and powering, respectively, while $\langle\cdot,\cdot\rangle_A$ denotes the Aitchison scalar product. Ultimately, this algebraic structure enables a mathematically principled extension of classical multivariate statistical tools to compositional data. See e.g. Aitchison, 1986, Van Den Boogaart and Tolosana-Delgado, 2013, Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado, 2015, Greenacre, 2018.

Despite their theoretical elegance, log-ratio methods face a major practical limitation: they require all components to be strictly positive. In real-world datasets, however, zeros occur frequently. They may represent structural absences (true zeros) or result from measurement limitations (rounded zeros, e.g. due to detection limits). The presence of zeros prevents the direct computation of logarithms and thus renders log-ratio transformations undefined. Naive substitution of zeros by small positive values can severely distort the geometry of the data and bias subsequent statistical analyses (Greenacre, 2021). Consequently, the *zero problem* has

[1] i.e. the interior of Δ_+^d.

become a central issue in modern compositional data analysis— quoting (Greenacre, 2021): "Zeros in compositional data are the Achilles heel of the logratio approach".

1.2. **Related works.** This zero-issue has motivated the development of further research, like specialized imputation procedures, zero-adjusted models, and alternative transformations. Among several alternative approaches, a notable one is the α-transformation proposed by Tsagris, Preston, and Wood, 2011; Tsagris, 2015; Tsagris, Preston, and Wood, 2016, which generalizes the centered log-ratio transformation by introducing a parameter $\alpha \in [0, 1]$. When $\alpha \to 0$, the transformation approaches the clr transformation, whereas for $\alpha > 0$, it can be applied to compositions containing zeros because it avoids taking logarithms. The α-transformation thus embeds the simplex into a Euclidean space through a smooth power transformation, allowing standard multivariate analyses to be performed while retaining a link with the log-ratio geometry. The choice of α can be guided by data-driven criteria, such as maximizing likelihood or optimizing classification performance.

A different line of work is proposed by Faugeras, 2026, 2025, which is based on the twin representation of the CoDa space, either as a projective space where CoDa are projective points which can be combined and studied via the exterior product, or as an affine space where CoDa are points in barycentric coordinates. These geometric viewpoints allow to develop a log-free divergence or distance for CoDa with zeroes, with corresponding variance/covariance matrices, enabling the decomposition of the variation of a sample among its pairs of components. These constructs open the way to the use of statistical methods that respect its underlying structure, without requiring any modification or imputation of the original data.

1.3. **Aims and scope.** The present work contributes to this line of research: to address the zero problem without relying on arbitrary substitution. It aims at overcoming some criticisms made to the above-mentioned works. Indeed, the α-transformation is not subcompositionally coherent, which has led some proponents of the CoDA school to reject it. From such a viewpoint, Subcompositional coherence is deemed essential for the interpretability and consistency of statistical analyses: it embodies the compositional principle that only ratios between observed parts should matter. A lack of coherence violates this key principle, meaning that analyses can become inconsistent when focusing on subcompositions. See e.g. Scealy and Welsh, 2014 for further debate.

On the other hand, the approaches of Faugeras, 2026 or Faugeras, 2025 are based on somehow unfamiliar geometric concepts and conceptually more involved mathematical constructions like formulas of displacement vectors in barycentric coordinates and Grassmann's exterior product and the Plücker embedding of projective spaces in the exterior algebra. These may present a significant learning curve for practitioners, making their adoption in practical data analysis more challenging and less accessible.

We thus ask whether it is possible to find CoDa representations

i) able to deal with zeroes,
ii) are easily interpretable in terms of ratios of components,
iii) which (at least partially) maintain subcompositional coherence,
iv) and are based on an easy framework of simply transforming the data, in a spirit similar to the $\alpha-$transformations.

The aim of this paper is to explore such possibilities.

1.4. **Outline.** We begin our investigations in Section 2 by inquiring the issue of zeros with ratios and their logarithms. The basic idea stems from the observation that a single (log)-ratio can be considered as an extended number and turned into a genuine number via a ratio-shaped homeomorphism σ, inducing a metric on these extended numbers. Section 3 apply this σ transform to multivariate CoDa, as these are represented by several ratios in the affine model. This yields a ratio-representation in the unit hypercube for CoDa with zeros in possibly all components except for a single reference one, with corresponding induced metric. An extension allows to handle CoDa with zeros in all components, except for zeros in some user-defined amalgamated parts. Pursuing the extension to its logical conclusion, Section 4 gives global representations of CoDa with zeros as points embedded in a subset of the Euclidean space. These coordinate representations interprets as ratios of "excesses" of the components. Section 5 study the compositional properties of the various proposed representations and their induced metrics. In particular, we contrast the pros and cons of the obtained local and global representations w.r.t. zero-coherence and subcompositional coherence and dominance. Section 6 illustrates the interest of the proposed representations by analyzing a real data set. We conclude our investigations in Section 7, with supplementary discussions in Section 8. Proofs are deferred to Appendix A.

2. Preliminaries: dealing with a (log-)ratio in a rational way

The foundational concept in CoDa analysis is the principle of scale invariance (Aitchison, 1986; Aitchison, 1992): information in a composition is contained in the *ratios* x_i/x_j, $i \neq j$, between its parts, not in their absolute values x_i, so that these ratios are the same whether the data are counts or are proportions, viz. $x_i/x_j = u_i/u_j$.

In view of this principle, any property, i.e. any mapping $f : \mathbb{R}_+^{d+1} \setminus \{\mathbf{0}\} \to \mathbb{R}$, is compositional iff it is scale invariant, i.e.

$$f(\lambda \mathbf{u}) = f(\mathbf{u}), \qquad \mathbf{u} \in \mathbb{R}_+^{d+1} \setminus \{\mathbf{0}\}, \lambda > 0.$$

In particular, coordinates mapping must be based on scale invariant functions. As pairwise ratios x_i/x_j are the minimal invariants, Aitchison hereby justify that CoDa analysis must be based on (and only on) ratios of components.

2.1. **Ratios and log-ratios of non-negative numbers as extended numbers.** When CoDa have zeros, infinities in Aitchison's log-ratios coordinates occur both from ratios and from logarithms. Indeed, ratios of proportions x_i/x_j, (equivalently, ratios of raw counts u_i/u_j) are obviously undefined in $\mathbb{R}$, if the denominator $x_j = 0$. If one sets $0/0 := 1$, then

$$\frac{x_i}{x_j} \begin{cases} \in [0, +\infty), & \text{if } x_j > 0 \\ = +\infty, & \text{if } x_j = 0, x_i > 0 \\ = 1, & \text{if } x_j = x_i = 0. \end{cases}$$

Hence, ratios of CoDa components become well-defined as elements of Aleksandrov's one-point compactification[2] $[0,\infty] := \mathbb{R}_+ \cup \{\infty\}$, obtained by adding a single improper element $\{\infty\}$ to $\mathbb{R}_+$.

Similarly, log-ratios (alr-transforms) $\ln(x_i/x_0)$ are undefined as real numbers, if either numerator x_i or denominator x_0 is zero, yielding either $-\infty$ or $+\infty$ (still with the convention that $0/0 = 1$). Since the $\ln : (0,\infty) \to \mathbb{R}$ function can be continuously extended to a function of extended numbers

$$\ln : [0,\infty] \to [-\infty,+\infty],$$

log-ratios of CoDa components become well-defined as elements of the two-points compactification $[-\infty,+\infty]$ (or affine extension)[3] of the real number line, obtained by adding the two improper elements $\{+\infty\}$ and $\{-\infty\}$ to $\mathbb{R}$, considered as an ordered set, see e.g. Bourbaki, 1960.

2.2. **Turning extended numbers into genuine real numbers.** The extended sets $[0,\infty]$ and $[-\infty,+\infty]$, containing infinities, are analytically and numerically awkward. Ratios, resp. log-ratios, can be transformed (and back) to more convenient (genuine) real numbers by any homeomorphism sending the intervals $[0,\infty]$, resp. $[-\infty,+\infty]$, to some closed bounded $[a,b]$, with $a < b \in \mathbb{R}$. Statistical literature traditionally focus on homeomorphisms $\sigma : [-\infty,+\infty] \to [0,1]$ obtained by cdfs of continuous distribution with positive density on $\mathbb{R}$, like the cdf Φ of a $\mathcal{N}(0,1)$ (inverse of probit), or of the logistic distribution function $x \mapsto 1/(1+e^{-x}) = 1/2 + 2^{-1}\tanh(x/2)$ (inverse of logit).

For our CoDa concerns, we will be interested in the *ratio-based* sigmoid function

$$\begin{aligned} \sigma : [-\infty,\infty] &\to [-1,1] \\ x &\mapsto \frac{x}{1+|x|}, \end{aligned} \tag{2}$$

whose inverse transform of is

$$\begin{aligned} \sigma^{-1} : [-1,1] &\to [-\infty,\infty] \\ y &\mapsto \frac{y}{1-|y|}. \end{aligned}$$

Its restriction[4] to $[0,\infty]$ writes as $\sigma(x) = x/(1+x)$ and is plotted in Figure 1. For x small, σ is approximately the identity (as $\sigma(0) = 0, \sigma'(0) = 1$), while it "squeezes" large values, which suggest it may be useful for handling skewed distributions of non-negative data.

[2]also called the projectively extended set of non-negative real numbers, or projective closure of $\mathbb{R}_+$. Adding "'points at infinity" is a classical approach to construct projective spaces from affine spaces, see e.g. Richter-Gebert, 2011.

[3]which is a different compactification of $\mathbb{R}$ than Aleksandrov's (projectively) extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. The latter identifies with the projective line $\mathbb{RP}^1$.

[4]We will confuse σ and its restriction to $[0,\infty]$. This should pose no ambiguities.

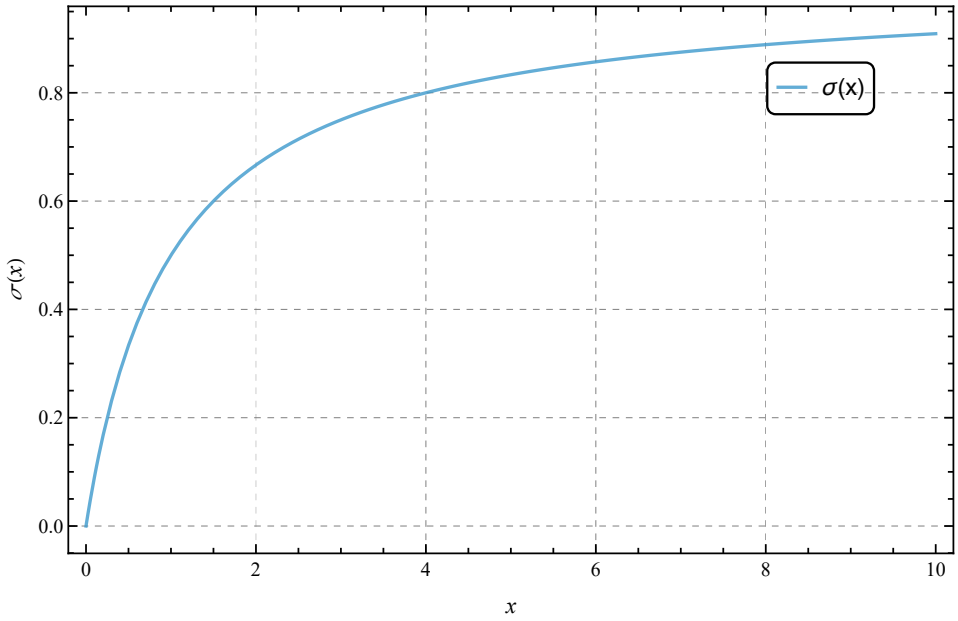

FIGURE 1. Ratio-based sigmoid functions $\sigma(x) = \frac{x}{1+x}$, restricted to $[0, \infty]$, realizing a homeomorphism of extended non-negative real numbers.

Indeed, when σ is applied to a ratio x_i/x_j, the extended number $x_i/x_j \in [0, \infty]$ turns[5] into

$$\sigma\left(\frac{x_i}{x_j}\right) = \frac{x_i/x_j}{1 + x_i/x_j} = \frac{x_i}{x_i + x_j} \in [0, 1], \tag{3}$$

a bounded regular number, on which numerical calculations can be performed. Interestingly, (3) maintains its ratio-form, as it writes as a ratio of linear forms, (that is homogeneous polynomials of degree one), involving only the components x_i and x_j. Hence, is a scale invariant *ratio of proportions*[6]. These $\sigma-$ratios are thus *compositional*, per Aitchison's requirement, and can serve, in principle, as bounded coordinate mappings on which further analysis can be based, as will be shown in the subsequent sections.

Similarly, when σ is applied to a log-ratio, the extended number $\ln(x_i/x_j) \in [-\infty, \infty]$ turns into a bounded, scale invariant quantity. However, for simplicity and maintaining the unity and focus of the paper, we will restrict in this paper only with ways to obtain surrogate *ratios* for CoDa with zeros. Section 8.2 quickly shows how the present σ-approach can be applied to obtain surrogates of *log-ratios* for CoDa with zeros. A more principled approach for obtaining surrogates of Aitchison's log-ratios (alr,clr,ilr) transforms for CoDa with zeros will be be dealt with in a subsequent paper, using a different generalization of the homeomorphism σ.

[5]still with the convention $0/0 := 1$.

[6]Of course, any function of ratios remain scale invariant. The point is that the function obtained remains bounded and interprets as a ratio of proportions.

2.3. **Distances on extended numbers.** Any distance on $[0,1]$ gives, via the homeomorphism σ, a distance on the extended set of numbers $[0,\infty]$. In particular, the usual (absolute value) distance $d(x,y) = |x-y|$ on $\mathbb{R}$ induces the following bounded distance for ratios:

$$d^\sigma(x,y) = |\sigma(x) - \sigma(y)| = \left| \frac{x}{1+x} - \frac{y}{1+y} \right|, \qquad x, y \in [0,\infty] \tag{4}$$

where x, y above stands for ratios, with the corresponding rule $\infty/\infty := 1$ for handling extended numbers. This endows the space of ratios with a metric structure.

3. Finite ratios for CoDa with zeros I: local representations

The previous section showed how to turn a single ratio as a finite (bounded) number by using a homeomorphism of ratio form, together with an induced distance. We now apply this to CoDa, i.e. to the multivariate case, as CoDa are represented by several ratios in their affine model.

3.1. **Application to ratios of CoDa.**

3.1.1. *Embedding of CoDa with zeros but with a common nonzero component.* For a general CoDa element $[\mathbf{x}]_+ \in \mathbb{P}^d_+$, (possibly with zeroes components), we can apply the σ transformation to the set of ratios

$$\pi_0([\mathbf{x}]_+) := \left(\frac{x_1}{x_0}, \dots, \frac{x_d}{x_0} \right) \in [0,\infty]^d. \tag{5}$$

That is to say, we define the transformation $\boldsymbol{\sigma}_0 := \boldsymbol{\sigma} \circ \pi_0$,

$$\begin{aligned} \boldsymbol{\sigma}_0 : \mathbb{P}^d_+ &\to \Sigma^d_0 := \boldsymbol{\sigma}_0(\mathbb{P}^d_+) \subset [0,1]^d \\ [\mathbf{x}]_+ &\mapsto \boldsymbol{\sigma}_0([\mathbf{x}]_+) = (\sigma(x_1/x_0), \dots, \sigma(x_d/x_0)) \end{aligned} \tag{6}$$

where

$$\sigma\left(\frac{x_i}{x_0}\right) = \frac{x_i/x_0}{1 + x_i/x_0} = \frac{x_i}{x_0 + x_i} \in [0,1], \quad i = 1, \dots, d,$$

and $\Sigma^d_0 := \boldsymbol{\sigma}_0(\mathbb{P}^d_+) \subset [0,1]^d$ denotes the image of the CoDa space by $\boldsymbol{\sigma}_0$.

Although $\boldsymbol{\sigma}_0$ is defined on the whole CoDa space $\mathbb{P}^d_+$, it is not injective everywhere. Let

$$\mathbb{U}^d_0 := \left\{ [\mathbf{x}]_+ \in \mathbb{P}^d_+ : x_0 = 0 \right\}$$

be the set of CoDa elements with zero first component. $\boldsymbol{\sigma}_0$ sends any element of $\mathbb{U}^d_0$ to the same point $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^d$ (still with the convention $0/0 = 0$ if $x_i = x_0 = 0$). Indeed, one has:

Lemma 3.1. *$\boldsymbol{\sigma}_0$ realizes an (injective) embedding (only) from $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$ into the unit cube, viz.*

$$\boldsymbol{\sigma}_0 : \mathbb{P}^d_+ \setminus \mathbb{U}^d_0 \hookrightarrow [0,1]^d,$$

and is bijective from $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$ onto $[0,1)^d$.

This lemma yields a ratio representation of any CoDa element $[\mathbf{x}]_+$ of $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$ as a point in the semi-open hypercube $[0,1)^d$, a subset of the Euclidean space $\mathbb{R}^d$. This representation is thus called local in the sense that it is not valid globally. The back-transformation is given by

$$\begin{aligned} \boldsymbol{\sigma}_0^{-1} : [0,1)^d &\to \mathbb{P}^d_+ \setminus \mathbb{U}^d_0 \\ \mathbf{z} = (z_1 \dots z_d) &\mapsto (x_0, x_1, \dots, x_d) \end{aligned}$$

with

$$\begin{cases} x_0 &= \frac{1}{1+\sum_{j=1}^d z_j/(1-z_j)} \\ x_i &= \frac{z_i/(1-z_i)}{1+\sum_{j=1}^d z_j/(1-z_j)}, \quad i=1,\dots,d. \end{cases} \tag{7}$$

Figure 2 illustrates the transformation for $d = 2$. The left figure shows the simplex Δ^2_+, with corresponding iso-lines of constant component for x_0 (red), x_1 (green), x_2 (blue). The right figure shows the image $\Sigma^2_0 = [0,1)^2 \cup \{(1,1)\}$ of the CoDa simplex space Δ^2_+ by $\boldsymbol{\sigma}_0$ as the unit square (light grey). The lines $x_1 = 0$ (dark green, bottom of left figure), resp. $x_2 = 0$ (dark blue, right side of the simplex on the left figure), located on the boundary of the simplex are sent to the vertical (dark green, right) line $z_1 = 0$, resp. horizontal $z_2 = 0$ (dark blue, right) line, in the unit square (right) by $\boldsymbol{\sigma}_0$ (except for the vertices with $x_0 = 0$, which are sent to $(1,1)$), while the iso-line $x_0 = 0$ (left, dark red) collapses to the single point $(1,1)$ (red, right). Lighter tones in the iso-lines stand for a corresponding higher constant component. For brevity, the legend only shows the constant x_0-lines.

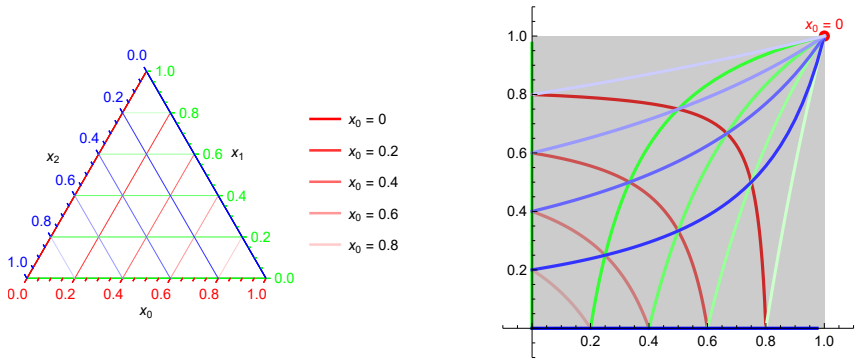

FIGURE 2. (Left): simplex representation of the CoDa space Δ^2_+, with iso-lines of constant component for x_0 (red), x_1 (green), x_2 (blue). (Right): local $\boldsymbol{\sigma}_0$-representation of $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$ as Σ^2_0, with corresponding images of the iso-lines.

The components of $x_i/(x_0 + x_i)$ of $\boldsymbol{\sigma}_0([\mathbf{x}]_+)$ measure the relative proportion of part i w.r.t. to parts $\{0, i\}$. They thus have a simple interpretation as measuring the relative strength of part i w.r.t. the base part 0 (amalgamated with i). $\boldsymbol{\sigma}_0$ can be thought as a sort of log-free version of the additive log-ratio w.r.t. part 0, enabling the treatment of zeros for parts $i > 0$.

3.1.2. *Ratio ℓ^p metric on $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$.* Applying an ℓ_p, $1 \leq p \leq \infty$ version of the distances d^σ of (4), yields the following extrinsic distance on $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$:

$$d_p^{\boldsymbol{\sigma}_0}([\mathbf{x}]_+, [\mathbf{y}]_+) := \|\boldsymbol{\sigma}_0([\mathbf{x}]_+) - \boldsymbol{\sigma}_0([\mathbf{y}]_+)\|_p \tag{8}$$
$$= \begin{cases} \left(\sum_{j=1}^d \left|\frac{x_i}{x_0+x_i} - \frac{y_i}{y_0+y_i}\right|^p\right)^{1/p}, & 1 \leq p < \infty \\ \max_{j=1,\dots d}\left|\frac{x_i}{x_0+x_i} - \frac{y_i}{y_0+y_i}\right|, & p = \infty \end{cases}$$

The latter is scale-invariant (being ratio-based), can handle zeroes for components $j = 1, \dots, d$. Any two CoDa points with null first component $x_0 = y_0 = 0$ are at zero distance. Hence, $d_p^{\boldsymbol{\sigma}_0}$ is only a pseudo-distance on $\mathbb{P}^d_+$. Restricted to elements of $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$, it induces a genuine bounded distance. (Note that it can compute a nonzero dissimilarity when only a single CoDa point is on $\mathbb{U}^d_0$ (i.e. when $[\mathbf{x}]_+, [\mathbf{y}]_+$ are s.t. $x_0 = 0$ and $y_0 \neq 0$).

Remark 1. *Notice that, for $1 \leq p < \infty$, $d_p^{\boldsymbol{\sigma}_0}$ writes as a weighted sum of determinants,*

$$d_p^{\boldsymbol{\sigma}_0}([\mathbf{x}]_+, [\mathbf{y}]_+) = \left(\sum_{j=1}^d \left|\frac{det\begin{vmatrix} x_j & y_j \\ x_0 & y_0 \end{vmatrix}}{(x_j + x_0)(y_j + y_0)}\right|^p\right)^{1/p}.$$

It thus interesting to note that we get a form similar to the barycentric divergence of Faugeras, 2026, and the exterior product distance of Faugeras, 2025, which were also based on sums of weighted determinants of order two.

Remark 2 (Comparison with affine distances)**.** *Applying directly the ℓ_p distance to ratios (5) gives, for $p < \infty$, the following metric on $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$,*

$$d_p^0([\mathbf{x}]_+, [\mathbf{y}]_+) := \|\pi_0([\mathbf{x}]_+) - \pi_0([\mathbf{y}]_+)\|_p = \left(\sum_{j=1}^d \left|\frac{x_i}{x_0} - \frac{y_i}{y_0}\right|^p\right)^{1/p}$$
$$= \left(\sum_{j=1}^d \left|\frac{det\begin{vmatrix} x_j & y_j \\ x_0 & y_0 \end{vmatrix}}{x_0 y_0}\right|^p\right)^{1/p},$$

and similarly for $p = \infty$. Compared to $d_p^{\boldsymbol{\sigma}_0}$, d_p^0 is infinite if $x_0 = 0, y_0 \neq 0$, or $y_0 = 0, x_0 \neq 0$, and is equal to the dimension d if both $x_0 = 0, y_0 = 0$ (for $p < \infty$, with the convention $0/0 = 1$). In turn, such metric an be turned into a bounded metric on $\mathbb{P}^d_+ \setminus \mathbb{U}^d_0$ by defining

$$\tilde{d}_p^0 := \frac{d_p^0}{1 + d_p^0} \leq 1.$$

The latter is

$$\tilde{d}_p^0([\mathbf{x}]_+, [\mathbf{y}]_+) = \sigma(\|\pi_0([\mathbf{x}]_+) - \pi_0([\mathbf{y}]_+)\|_p)$$

whereas (8) is

$$d_p^{\boldsymbol{\sigma}_0}([\mathbf{x}]_+, [\mathbf{y}]_+) = \|\boldsymbol{\sigma} \circ \pi_0([\mathbf{x}]_+) - \boldsymbol{\sigma} \circ \pi_0([\mathbf{y}]_+)\|_p.$$

In other words, the order in the composition is reversed.

3.1.3. *Towards statistical applications.* The statistical interest of the ratio representation $\boldsymbol{\sigma}_0$ is for compositional datasets having a component always non-zero, i.e. s.t. $x_0 \neq 0$ for all data points: $\boldsymbol{\sigma}_0$ gives a representation of CoDa elements of $\mathbb{P}_+^d \setminus \mathbb{U}_0^d$ as a familiar point/vector of a subset of $[0,1]^d$. One can then apply the distances $d_p^{\boldsymbol{\sigma}_0}$ which is able to handle zeroes in other components. In turn, one can define notions of center (Fréchet means), variance matrices, etc., as in Faugeras, 2026, Faugeras, 2025. More generally, since elements $[\mathbf{x}]_+$ of $\mathbb{P}_+^d \setminus \mathbb{U}_0^d$ are embedded via $\boldsymbol{\sigma}_0$ as points in the subset $\Sigma_0^d \setminus \{\mathbf{1}\}$ which is itself a subset of the Euclidean space $(\mathbb{R}^d, \langle .|. \rangle)$, one can use classical multivariate techniques, for CoDa with zeros outside of a common component. See Section 6 for a numerical illustration based on linear regression on a real dataset.

3.2. **Generalization to more than a single reference component.** The approach of Section 3.1 allowed to reduce CoDa datasets where a component is always non-zero. In cases where no single component is non-zero for the whole dataset, one can look for ratio embeddings based on more than component in the denominator. That is to say, for $1 \leq k < d$, one consider the set of ratios

$$\pi_{0\ldots k}([\mathbf{x}]_+) := \left(\frac{x_1}{\sum_{j=0}^k x_j}, \ldots, \frac{x_d}{\sum_{j=0}^k x_j} \right) \in [0,\infty]^d, \tag{9}$$

(which corresponds to representing the ray $[\mathbf{x}]_+$ by its intersection point on the hyperplane $\sum_{j=0}^k x_j = 1$), and apply the σ transform component-wise. This results in a mapping

$$\begin{aligned} &\boldsymbol{\sigma}_{0\ldots k} : \mathbb{P}_+^d \to \Sigma_{0\ldots k}^d := \boldsymbol{\sigma}_{0\ldots k}(\mathbb{P}_+^d) \subset [0,1]^d \\ &[\mathbf{x}]_+ \mapsto \boldsymbol{\sigma}_{0\ldots k}([\mathbf{x}]_+) = \left(\sigma\left(\frac{x_1}{\sum_{j=0}^k x_j} \right), \ldots, \sigma\left(\frac{x_d}{\sum_{j=0}^k x_j} \right) \right). \end{aligned} \tag{10}$$

The components $(z_i)_{i=1,\ldots,d}$ of $\boldsymbol{\sigma}_{0\ldots k}([\mathbf{x}]_+)$ are given by

$$z_i := (\boldsymbol{\sigma}_{0\ldots k}([\mathbf{x}]_+))_i = \frac{x_i}{\sum_{j=0}^k x_j + x_i}, \quad i = 1, \ldots, d. \tag{11}$$

The denominator in (11) interprets as an amalgamation of each part $i = 1, \ldots d$ with the parts $\{0, \ldots, k\}$, counting twice component x_i when $i \leq k$. However, this approach is different from amalgamation of components with zeroes into a single component, which, by reducing the dimension, compresses the original data and distorts its informational content. Here, $\boldsymbol{\sigma}_{0\ldots k}$ of (10) (and $\boldsymbol{\sigma}_0$ of (6)) keep the original dimension d of the data. (See Faugeras, 2026 Section 7.1–7.2 for a principled approach to amalgamation.)

Remark 3 (A rescaled variant)**.** *Since for $1 \leq i \leq k$, x_i is counted twice in the denominator of (11), the components z_i of (10) satisfy $0 \leq z_i \leq 1/2$, with the upper bound attained for $\sum_{j \neq i, j \leq k} x_j = 0$, while for $k < i \leq d$, $0 \leq z_i \leq 1$, with the upper bound attained for $\sum_{j \leq k} x_j = 0$. In other words, the image $\Sigma_{0\ldots k}^d$ is (a subset of) the hyper rectangle $[0,1/2]^k \times [0,1]^{d-k}$. In order to have all components scaled to the same unit interval, one can consider the following variant which sends $[\mathbf{x}]_+$ to the point of coordinates*

$$\left(\frac{2x_1}{\sum_{j=0}^k x_j + x_1}, \ldots, \frac{2x_k}{\sum_{j=0}^k x_j + x_k}, \frac{x_{k+1}}{\sum_{j=0}^k x_j + x_{k+1}}, \ldots, \frac{x_d}{\sum_{j=0}^k x_j + x_d} \right). \tag{12}$$

This corresponds to transforming components $1,\dots,k$ *of (9) by* 2σ*, and components* $k+1,\dots,d$ *by* σ*.*

The properties of (10) or the variant (12) are similar to those of Lemma 3.1 and are thus omitted. The corresponding ℓ_p distances $d_p^{\boldsymbol{\sigma}_{0\dots k}}$ can be defined accordingly, as in (8), yielding a pseudo-metric on the whole CoDa space, and a genuine metric on the CoDa $[\mathbf{x}]_+$ s.t. $\sum_{j=0}^k x_j \neq 0$.

For statistical analysis purposes on a given dataset, one can take k as the minimal number of components so that the sum $\sum_{j=0}^k x_j > 0$ for all data points. For suitable k and ordering of the components, this covers almost all CoDa sets, to the exception of the most severe ones with zeros in every components. This distance can thus be applied on almost all CoDa sets with zeros, enabling statistical analysis based on metrical notions.

3.3. **Towards a global affine representation for CoDa with zeroes.** Going all the way in the ratio normalization (9) with all $d+1$ components in the denominator will lead to the approach of the next Section, upon realizing that the number 1 is itself a sum of CoDa components, when these are represented in the simplex.

Indeed, dividing x_i by the sum of all $d+1$ components $\sum_{j=0}^d x_j$ in (9) gives as affine representation of a CoDa $[\mathbf{x}]_+$ the intersection point of the ray $[\mathbf{x}]_+$ with the ($d-$dimensional) hyperplane $\sum_{j=0}^d x_j = 1$, i.e.

$$\pi_{0,\dots,d}([\mathbf{x}]_+) := \left(\frac{x_1}{\sum_{j=0}^d x_j},\dots,\frac{x_d}{\sum_{j=0}^d x_j}\right).$$

Thus, up to the omission of the first component $x_0/\sum_{j=0}^d x_j$, (which was done on purpose in order to match the dimension of $\mathbb{P}_+^d$), $\pi_{0,\dots,d}([\mathbf{x}]_+)$ corresponds to the classical simplex representation (1)

$$\pi_{0,\dots,d}([\mathbf{x}]_+) = \frac{\mathbf{x}}{|||\mathbf{x}||_1} = \mathcal{C}(\mathbf{x}) \in \Delta_+^d,$$

obtained by the closure operation in the CoDa literature. The interest of such affine representation is that it is global, compared to, say, π_0, which is undefined (as real numbers) on $\mathbb{U}_0^d$. We thus consider thereafter directly a CoDa point as a normalized element of the simplex Δ_+^d.

4. Finite ratios for CoDa with zeros II: global representations

4.1. **Application of σ to simplex components of CoDa.** For a CoDa element represented as an element $\mathbf{x}$ of the simplex Δ_+^d, it is noteworthy that the σ transformation applied directly to d components[7] of $\mathbf{x}$, i.e. setting

$$\boldsymbol{\sigma}_*(\mathbf{x}) := (\sigma(x_1),\dots,\sigma(x_d)) \tag{13}$$

also yields a ratio-based scale-invariant representation. Indeed, since $\mathbf{x} \in \Delta_+^d$, $\sum_{j=0}^d x_j = 1$, one has that

$$\sigma(x_i) = \frac{x_i}{1+x_i} = \frac{x_i}{\sum_{j=0}^d x_j + x_i}, \quad i = 1,\dots,d,$$

[7]W.l.o.g. we take parts $1,\dots,d$. Note that $\boldsymbol{\sigma}_*$ thus depends on the omitted part 0. We suppressed this dependence in order to have a less cumbersome notation.

and is thus scale invariant. Therefore, $\boldsymbol{\sigma}_* : \Delta_+^d \to \Sigma_*^d := \boldsymbol{\sigma}_*(\Delta_+^d)$ extends to a function $\boldsymbol{\sigma}_* : \mathbb{P}_+^d \to \Sigma_*^d$ on the whole CoDa space by setting

$$\boldsymbol{\sigma}_*([\mathbf{x}]_+) := \boldsymbol{\sigma}_*(\mathcal{C}(\mathbf{x})) = \boldsymbol{\sigma}_*\left(\frac{\mathbf{x}}{\|\mathbf{x}\|_1}\right) = \left(\frac{x_1}{\sum_{j=0}^d x_j + x_1}, \ldots, \frac{x_d}{\sum_{j=0}^d x_j + x_d}\right).$$

Moreover, one has the analogue of lemma 3.1:

Lemma 4.1. *$\boldsymbol{\sigma}_*$ is now injective on $\mathbb{P}_+^d$ and thus realizes an embedding of the whole CoDa space*

$$\boldsymbol{\sigma}_* : \mathbb{P}_+^d \hookrightarrow \Sigma_*^d \subset [0, 1/2]^d$$

into a subset of an Euclidean space. $\boldsymbol{\sigma}_*$ thus becomes a bijection onto its image, with inverse transformation $\boldsymbol{\sigma}_*^{-1} : \Sigma_*^d \to \Delta_+^d$ given by

$$\boldsymbol{\sigma}_*^{-1}(\mathbf{z}) = \left(1 - \sum_{j=1}^d \frac{z_j}{1 - z_j}, \frac{z_1}{1 - z_1}, \ldots, \frac{z_d}{1 - z_d}\right),$$

for $\mathbf{z} = (z_1, \ldots, z_d) \in \Sigma_*^d$.

Since $0 \leq x_i/(1 + x_i) \leq 1/2$, for $0 \leq x_i \leq 1$, Σ_*^d is subset of the hyperrectangle $[0, 1/2]^d$, and is characterized analytically as

$$\Sigma_*^d = \left\{\mathbf{z} \in \mathbb{R}^d : \sum_{j=1}^d \frac{z_j}{1 - z_j} \leq 1, z_i \geq 0, \quad i = 1, \ldots, d\right\}$$

Figure 3 shows the region obtained, for $d = 2$ (3 components), viz. $\Sigma_*^2 = \{(z_1, z_2) : z_1/(1 - z_1) + z_2/(1 - z_2) \leq 1, z_1, z_2 \geq 0\}$ and gives a comparison with the spherical representation $\{(z_1, z_2) : z_1^2 + z_2^2 \leq 1, z_1, z_2 \geq 0\}$. Also drawn are the images by $\boldsymbol{\sigma}_*$ of the iso-lines of constant x_0 (red), x_1 (green), x_2 (Blue) component of the simplex of Figure 2. The color code is the same as Figure 2, for easing comparisons.

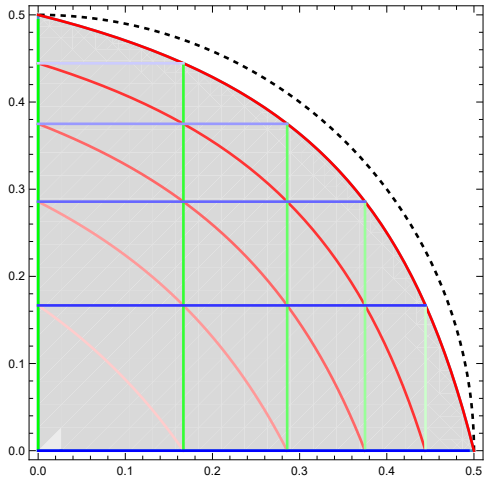

FIGURE 3. Region Plot of Σ^2_* (Gray), with images by $\boldsymbol{\sigma}_*$ of the iso-lines for x_0 (red), x_1 (green), x_2 (Blue) of the simplex of Figure 2, for $d = 2$ (3 components). The non-negative circle of radius $1/2$ (dashed black) corresponding to a spherical representation of CoDa is drawn for comparison.

Applying an ℓ_p, $1 \leq p \leq \infty$ distance on the $\boldsymbol{\sigma}_*$ representations of the CoDa elements, yields the following global distance on $\mathbb{P}^d_+$ (i.e. scale invariant):

$$d_p^{\boldsymbol{\sigma}_*}([\mathbf{x}]_+, [\mathbf{y}]_+) = \|\boldsymbol{\sigma}_*(\mathcal{C}(\mathbf{x})) - \boldsymbol{\sigma}_*(\mathcal{C}(\mathbf{y}))\|_p \tag{14}$$

For $\mathbf{x}, \mathbf{y} \in \Delta^+_d$, the latter writes as

$$d_p^{\boldsymbol{\sigma}_*}([\mathbf{x}]_+, [\mathbf{y}]_+) = \begin{cases} \left(\sum_{i=1}^d \left|\frac{x_i - y_i}{(1+x_i)(1+y_i)}\right|^p\right)^{1/p} & 1 \leq p < \infty \\ \max_{i=1,\dots d} \frac{|x_i - y_i|}{(1+x_i)(1+y_i)} & p = \infty \end{cases}$$

4.2. **A second global variant: ratio of excesses.** Realizing fully the trivial observation that for $\mathbf{x} \in \Delta^d_+$, the constant 1 is in fact the sum of simplex components and thus that ratios of $1, x_0, \dots, x_d$ give ratios of CoDa components $(x_0, \dots, x_d)$, suggest the following transformation to deal globally for CoDa with zeros:

$$\begin{aligned} \pi_0^+ : \Delta^d_+ &\to \Sigma^d_{0,+} := \pi_0^+(\mathbb{P}^d_+) \subset [2^{-1}, 2]^d \\ [\mathbf{x}]_+ &\mapsto \left(\frac{1+x_1}{1+x_0}, \dots, \frac{1+x_d}{1+x_0}\right). \end{aligned}$$

The latter is thus the analogue of π_0 of (5), but "starting" from 1 in both the numerator and denominator. It is inspired by the starting trick in exploratory data analysis, see Tukey, 1977, Mosteller and Tukey, 1977 p. 83, where e.g. log transforms are performed on shifted data $x \mapsto \ln(x + c)$ in case these contain non-positive values, with suitably chosen starting value c.

Such transformation extends to a scale-invariant transformation (which we also denote by π_0^+) on the absolute amounts $\mathbf{u} \in \mathbb{R}_+^{d+1} \setminus \{\mathbf{0}\}$, as

$$\begin{aligned} \pi_0^+ : \mathbb{R}_+^{d+1} \setminus \{\mathbf{0}\} &\to \Sigma_{0,+}^d \subset [2^{-1}, 2]^d \\ \mathbf{u} &\mapsto \left(\frac{\|\mathbf{u}\|_1 + u_1}{\|\mathbf{u}\|_1 + u_0}, \ldots, \frac{\|\mathbf{u}\|_1 + u_d}{\|\mathbf{u}\|_1 + u_0} \right). \end{aligned} \tag{15}$$

Indeed, since $\mathbf{u} \geq \mathbf{0}$ and $\mathbf{u} \neq 0$, the denominators in (15) never vanish and π_0^+ is well-defined. As a ratio of linear forms, it satisfies scale invariance, viz.

$$\pi_0^+(\lambda \mathbf{u}) = \pi_0^+(\mathbf{u}), \quad \lambda > 0.$$

In other words, π_0^+ is a well-defined transformation on the CoDa space of equivalence classes $[\mathbf{x}]_+ \in \mathbb{P}_+^d$, see Figure 4 .

$$\begin{array}{ccc} \mathbb{R}^{d+1} \setminus \{\mathbf{0}\} & \xrightarrow{\quad \mathcal{C} \quad} & \Delta_+^d \\ \mathbf{u} = (u_0, \ldots, u_d) & & \mathbf{x} = \frac{\mathbf{u}}{\|\mathbf{u}\|_1} \\ \downarrow \pi_0^+ & & \downarrow \pi_0^+ \\ \Sigma_{0,+}^d & & \Sigma_{0,+}^d \\ \pi_0^+(\mathbf{u}) := \left(\frac{\|\mathbf{u}\|_1 + u_1}{\|\mathbf{u}\|_1 + u_0}, \ldots, \frac{\|\mathbf{u}\|_1 + u_d}{\|\mathbf{u}\|_1 + u_0} \right) & \xleftrightarrow{\text{id}} & \pi_0^+(\mathbf{x}) := \left(\frac{1+x_1}{1+x_0}, \ldots, \frac{1+x_d}{1+x_0} \right) \end{array}$$

FIGURE 4. Lifting of π_0^+ by scale invariance to the CoDa space $\mathbb{P}_+^d$: π_0^+ applied to raw counts $\mathbf{u}$ gives an identical transformation as π_0^+ applied to closed elements $\mathbf{x}$ of the simplex Δ_+^d.

In fact, one has the analogue of the previous Lemmas 3.1 and 4.1

Lemma 4.2. *$\pi_0^+ : \mathbb{P}_+^d \to \Sigma_{0,+}^d$ is bijective on the whole CoDa space, with inverse transformation given by the following fractional-linear transformation (Möbius transform):*

$$\begin{aligned} (\pi_0^+)^{-1} : \Sigma_{0,+}^d &\to \mathbb{P}_+^d \\ (z_1, \ldots z_d) &\mapsto [\mathbf{x}]_+, \end{aligned}$$

where $\mathbf{x} \in \Delta_+^d$ is given by

$$\begin{aligned} x_0 &= \frac{d + 1 - \sum_{j=1}^d z_j}{1 + \sum_{j=1}^d z_j}, \\ x_i &= z_i(1 + x_0) - 1 = \frac{(d+2) z_i}{1 + \sum_{j=1}^d z_j} - 1, \quad i = 1, \ldots, d. \end{aligned}$$

The characterization of the simplex Δ_+^d by $x_i \geq 0$, $i = 1, \ldots, d$ and $\sum_{j=1}^d x_j \leq 1$ gives an analytical description of $\Sigma_{0,+}^d$ as a polyhedral region. Figure 5 shows $\Sigma_{0,+}^d$ for $d = 2$ in the (z_1, z_2) plane. $\Sigma_{0,+}^2$ (grey triangle) is analytically characterized by $z_1 + z_2 \leq 3$, $1 - 3z_1 + z_2 \leq 0$, $1 - 3z_2 + z_1 \leq 0$.

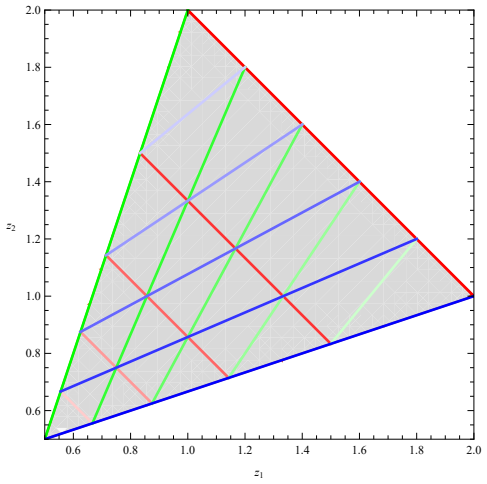

FIGURE 5. Region plot of $\Sigma^2_{0,+}$ (grey), for $d = 2$, with images by π_0^+ of the iso-lines for x_0 (red), x_1 (green), x_2 (blue) of the simplex of Figure 2, for $d = 2$ (3 components).

The components z_i of the π_0^+-transformed CoDa elements have an interesting interpretation as measuring the ratio of the "excess" $1 + x_i$ of the $i-$th component x_i from the whole composition 1, to the corresponding excess of the $0-$th component x_0 w.r.t the whole composition 1. It thus measures the relative strength of the x_i component w.r.t. to x_0, relatively to the total.

As previously, one can define a corresponding bounded distance on the whole CoDa space, by applying an ℓ_p norm to the difference of the π_0^+ representations of the CoDa points.

$$d_p^{\pi_0^+}([\mathbf{x}]_+, [\mathbf{y}]_+) = \|\pi_0^+([\mathbf{x}]_+) - \pi_0^+([\mathbf{y}]_+)\|_p \tag{16}$$

However, the latter distance is not zero-coherent. On the other hand, since the components of $\pi_0^+([\mathbf{x}]_+)$ are in $[1/2, 2]$, which does not contain zero, one can apply the ln function to obtain a representation in $[-\ln 2, \ln 2]$, i.e. in a symmetrical interval.

5. Compositional properties of the proposed representations

We discuss in this section how the previously obtained local and global representations behave w.r.t. desirable compositional properties, in particular compositional coherence.

5.1. **Zero-coherence.** Since a composition can always be seen as a subcomposition of a larger composition with added zeros, it is natural to expect that the way one measures the distance between compositions is coherent with the operation of adding zeros. This property can be called the principle of zero-coherence. See Faugeras, 2026 for further discussion. It is worth stressing at this point that Aitchison's distance does not satisfy this principle.

Since $\sigma(0) = 0$, one has that any representation based on σ permutes with the concatenation–with–zeros operation. In other words, one has that $\boldsymbol{\sigma}_0([\mathbf{x} : \mathbf{0}]_+) =$

$(\boldsymbol{\sigma}_0([\mathbf{x}]_+) : \mathbf{0})$ and $\boldsymbol{\sigma}_*([\mathbf{x} : \mathbf{0}]_+) = (\boldsymbol{\sigma}_*([\mathbf{x}]_+) : \mathbf{0})$. Metrically, this translates into the following proposition.

Proposition 5.1 (zero-coherence)**.** *i) If $[\mathbf{x}]_+, [\mathbf{y}]_+ \in \mathbb{P}^d \setminus \mathbb{U}_0^d$ are seen as subcompositions of a larger composition with zeros $[\mathbf{x} : \mathbf{0}]_+, [\mathbf{y} : \mathbf{0}]_+ \in \mathbb{P}_+^{d+k} \setminus \mathbb{U}_0^{d+k}$, where $\mathbf{0} \in \mathbb{R}^k$, for some integer $k > 0$, and where $\mathbf{x} : \mathbf{0}$ stands for the concatenation of $\mathbf{x}$ and $\mathbf{0}$, then the induced local distances $d_p^{\boldsymbol{\sigma}_0}$ of (8) and $d_p^{\boldsymbol{\sigma}_{0\dots k}}$ of (10) satisfy*

$$d_p^{\boldsymbol{\sigma}_0}([\mathbf{x} : \mathbf{0}]_+, [\mathbf{y} : \mathbf{0}]_+) = d_p^{\boldsymbol{\sigma}_0}([\mathbf{x}]_+, [\mathbf{y}]_+)$$
$$d_p^{\boldsymbol{\sigma}_{0\dots k}}([\mathbf{x} : \mathbf{0}]_+, [\mathbf{y} : \mathbf{0}]_+) = d_p^{\boldsymbol{\sigma}_{0\dots k}}([\mathbf{x}]_+, [\mathbf{y}]_+).$$

ii) Similarly for the global distance $d_p^{\boldsymbol{\sigma}_*}$ of (14), one has

$$d_p^{\boldsymbol{\sigma}_*}([\mathbf{x} : \mathbf{0}]_+, [\mathbf{y} : \mathbf{0}]_+) = d_p^{\boldsymbol{\sigma}_*}([\mathbf{x}]_+, [\mathbf{y}]_+),$$

for $[\mathbf{x}]_+, [\mathbf{y}]_+ \in \mathbb{P}^d$.

Regarding $d_p^{\pi_0^+}$, we note that it does not satisfy zero-coherence.

5.2. **Partial subcompositional coherence.** Subcompositional coherence is often given in the literature a vague, verbal definition, such as "There should be agreement between statements on a subcomposition whether they are obtained from analyzing the full composition or just the subcomposition." (Scealy and Welsh, 2014). A further distinction is often made between subcompositional coherence per se, and subcompositional dominance, i.e. "that distances between subcompositions can not be larger than distance between full subcompositions". Mathematically, the subcompositional coherence principle is formalized by requiring that the relevant statistical analysis or operation commutes with the operation of taking a subcomposition (Gajer and Ravel, 2025). See also the discussion in Faugeras, 2026 Section 3.2 and the references therein.

For discussing how the proposed representations behave w.r.t. subcompositional coherence, we first introduce some precise notation. For some subset $J = \{i_1, \dots, i_m\} \subset \{0, \dots, d\}$ of indices of cardinality $m := \mathrm{card}(J) \geq 2$, with $0 \leq i_1 < \dots < i_m \leq d$, denote by

$$\mathcal{S}_J : \mathbb{R}^{d+1} \to \mathbb{R}^m$$
$$(x_0, x_1, \dots, x_d) \mapsto (x_{i_1}, \dots, x_{i_m}) = (x_i)_{i \in J}$$

the linear sub-setting operation keeping the components of $\mathbf{x} = (x_i)_{i=0,\dots,d}$ with $i \in J$. If $\|\mathcal{S}_J(\mathbf{x})\|_1 \neq 0$, (i.e. if the selected components have positive total mass), $\mathcal{S}_J$ induces, by closure, the (fractionally linear) subcomposition operation $\mathcal{S}_J^\Delta$ on the CoDa simplex, viz.

$$\mathcal{S}_J^\Delta : \Delta_+^d \to \Delta_+^{k-1}$$
$$(x_0, x_1, \dots, x_d) \mapsto \mathcal{C} \circ \mathcal{S}_J(\mathbf{x}) := \frac{\mathcal{S}_J(\mathbf{x})}{\|\mathcal{S}_J(\mathbf{x})\|_1} = \frac{\mathcal{S}_J(\mathbf{x})}{\sum_{i \in J} x_i}.$$

In words, tshe subcomposition operator $\mathcal{S}_J^\Delta$ is defined by selecting the relevant components and re-closing them.

Now, let $I := \{i_1, \dots, i_k\} \subset \{1, \dots, d\}$ be a subset of indices of cardinality $k := \mathrm{card}(I)$, with part 0 excluded, i.e. with $1 \leq i_1 < \dots < i_k \leq d$, and denote by

$\mathcal{S}_I$, resp. $\mathcal{S}_{\{0\}\cup I}$, the sub-setting operations to I, resp. to $\{0\}\cup I$, viz.

$$\mathcal{S}_I : \mathbb{R}^{d+1} \to \mathbb{R}^k$$
$$(x_0, x_1, \ldots, x_d) \mapsto (x_{i_1}, \ldots, x_{i_k})$$

and

$$\mathcal{S}_{\{0\}\cup I} : \mathbb{R}^{d+1} \to \mathbb{R}^{k+1}$$
$$(x_0, x_1, \ldots, x_d) \mapsto (x_0, x_{i_1}, \ldots, x_{i_k}).$$

For CoDa elements with $x_0 \neq 0$, one denotes similarly by $\mathcal{S}_I^{\Delta} : \Delta_+^d \to \Delta_+^{k-1}$, resp. $\mathcal{S}_{\{0\}\cup I}^{\Delta} : \Delta_+^d \to \Delta_+^k$, the corresponding subcompositions operations with parts $i \in I$, resp. with parts $i = 0$ and $i \in I$, retained.

The ratio form of σ leads to the following subcompositional coherence property of $\boldsymbol{\sigma}_0$, for subcompositions containing part 0:

Proposition 5.2. *i) The $\boldsymbol{\sigma}_0$ representation (6) is subcompositionally-coherent on $\mathbb{P}_+^d \setminus \mathbb{U}_0^d$, for subcompositions $\mathcal{S}_{\{0\}\cup I}^{\Delta}$ containing part 0, where $I \subset \{1, \ldots, d\}$. In other words, the following diagram commutes*

$$\begin{array}{ccccc} \mathbb{R}_+^{d+1} \setminus \{\mathbf{0}\} & \xrightarrow{\mathcal{C}} & \Delta_+^d & \xrightarrow{\boldsymbol{\sigma}_0} & [0,1)^d \\ \downarrow{\scriptstyle \mathcal{S}_{\{0\}\cup I}} & & \downarrow{\scriptstyle \mathcal{S}_{\{0\}\cup I}^{\Delta}} & & \downarrow{\scriptstyle \mathcal{S}_I} \\ \mathbb{R}_+^{k+1} \setminus \{\mathbf{0}\} & \xrightarrow{\mathcal{C}} & \Delta_+^k & \xrightarrow{\boldsymbol{\sigma}_0} & [0,1)^k \end{array}$$

FIGURE 6. Subcompositional coherence for $\boldsymbol{\sigma}_0$.

ii) In particular, for the induced distances $d_p^{\boldsymbol{\sigma}_0}$ of (8), one has the corresponding subcompositional dominance:

$$d_p^{\boldsymbol{\sigma}_0}([\mathcal{S}_{\{0\}\cup I}^{\Delta}(\mathbf{x})]_+, [\mathcal{S}_{\{0\}\cup I}^{\Delta}(\mathbf{y})]_+) \leq d_p^{\boldsymbol{\sigma}_0}([\mathbf{x}]_+, [\mathbf{y}]_+),$$

for all $[\mathbf{x}]_+, [\mathbf{y}]_+ \in \mathbb{P}_+^d \setminus \mathbb{U}_0^d$.

Hence, Proposition 5.2 i) means that one obtains the same $\boldsymbol{\sigma}_0$ representation of the subcomposition $\mathcal{S}_{\{0\}\cup I}^{\Delta}(\mathbf{x}) \in \Delta_+^k$ (equivalently, of the sub-vector $\mathcal{S}_{\{0\}\cup I}(\mathbf{u}) \in \mathbb{R}_+^{k+1} \setminus \{\mathbf{0}\}$ of raw counts) as one would get by simply selecting, via $\mathcal{S}_I$, the components in I of the $\boldsymbol{\sigma}_0$-representation of the full composition $\mathbf{x} \in \Delta_+^d$ (equivalently of the full vector of raw amounts $\mathbf{u} \in \mathbb{R}_+^{d+1} \setminus \{\mathbf{0}\}$.

For the transformation $\boldsymbol{\sigma}_{0\ldots k}$ of (10), one has similar restricted subcompositional coherence properties, for subcompositions keeping the parts $\{0, \ldots, k\}$, i.e. for $\mathcal{S}_{\{0,\ldots,k\}\cup I}^{\Delta}$, with $I \subset \{k+1, \ldots, d\}$. The statement and proof are omitted. For comparison, neither $\boldsymbol{\sigma}^*$, nor Tsagris' α-transform do satisfy subcompositional coherence in the above sense.

Remark 4. *Proposition 5.2 shows that σ, applied to the ratios π_0, is subcompositionally coherent for subcompositions including part 0. To the contrary, σ applied*

directly to the components of $\mathbf{x} \in \Delta_+^d$ *is not subcompositionally coherent but satisfy a monotonicity property. Indeed, since* $\|\mathcal{S}_J(\mathbf{x})\|_1 \leq \|\mathbf{x}\|_1$*, one has that*

$$\sigma\left(\frac{x_i}{\|\mathcal{S}_J(\mathbf{x})\|_1}\right) = \frac{x_i}{\|\mathcal{S}_J(\mathbf{x})\|_1 + x_i} \geq \frac{x_i}{1+x_i} = \sigma(x_i), \quad i \in J.$$

This reflects the affine nature of CoDa as weighted points in barycentric coordinates, i.e. the fact that amalgamation and subcomposition are connected, as taking a subcomposition involves standardization by the amalgamation total $\|\mathcal{S}_J(\mathbf{x})\|_1$ *of the sub-vector* $\mathcal{S}_J(\mathbf{x})$*. See Faugeras, 2025, in particular his Section 3.2.*

5.3. **Discussion and Table.** A summary of the properties of the proposed representations and corresponding induced distances is given in Table 1. For comparison purposes, we also include Aitchison's log-ratios and distance d_A, together with Tsagris' $\alpha-$ transformation with corresponding distance $d_{T,\alpha}$.

The table shows that one has a "no free lunch" scenario. No single transformation satisfy all desirable properties: ability to handle zeros, zero-coherence, and subcompositional coherence. In addition, it appears there is some kind of interesting trade-off between the breadth of handling zeros and the subcompositional coherence property. At one extreme, log-ratios are fully coherent but unable to intrinsically handle zeros. At the other extreme, global representations, which can handle zeros in any components, like Tsagris' α or our $\boldsymbol{\sigma}^*$ or π_0^+, are not subcompositionally coherent. As a middle-ground, the local representations $\boldsymbol{\sigma}_0$ and $\boldsymbol{\sigma}_{0\ldots k}$ allow for restricted treatment of zeros and partial coherence. $\boldsymbol{\sigma}_{0\ldots k}$ allows for treating more zeros but less subcompositions than $\boldsymbol{\sigma}_0$, as was proved in the referenced Propositions and Lemmas. The color code of Table 1 highlights these features.

Transformation	Induced distance	Zero-handling	Zero-coherence	Subcomp. coherence
Aitchison's log-ratios	d_A	No	No	Yes
Tsagris' α-transforms	$d_{T,\alpha}$	Yes	Yes	No
$\boldsymbol{\sigma}_0$	$d_p^{\boldsymbol{\sigma}_0}$	Yes, local Lemma 3.1	Yes Proposition 5.1 i)	Yes, partial Proposition 5.2
$\boldsymbol{\sigma}_{0\ldots k}$	$d_p^{\boldsymbol{\sigma}_{0\ldots k}}$	Yes, quasi-global	Yes	Yes, restricted partial
$\boldsymbol{\sigma}_*$	$d_p^{\boldsymbol{\sigma}_*}$	Yes, global Lemma 4.1	Yes Proposition 5.1 ii)	No
π_0^+	$d_p^{\pi_0^+}$	Yes, global Lemma 4.2	No	No

TABLE 1. Comparison of transformations and their properties.

6. Statistical illustration: predicting the refractive index in glass data by subcompositionally coherent methods

6.1. **Data and methodology.** The glass dataset from UCI repository[8] is a classical data frame with 214 observation containing examples of the chemical composition (Na, Mg, Al, Si, K, Ca, Ba, Fe) of different types of glass. The problem is to forecast the Refractive Index (RI) (the ability to bend light sharply) on basis of the chemical analysis. (A higher refractive index allows lenses to be thinner, lighter, and more powerful). It contains 392 zeros in components Mg (42 zeros or 19.6% of the sample size), K (30 or 14%), Ba (176 or 82.2%), Fe (144 or 67.3%). Since glass is primarily made of Silicium (the formative oxide), with added components (fluxes, stabilizers and other modifier oxides), the Si component is dominant and never zero. This suggests to select Si as reference component (i.e. component x_0 in the notation of the present paper), and express relative variation of other components x_i w.r.t. Si.

Because the interest is in identifying and quantifying which added oxides has an impact on the refractive index, it makes sense to consider only subcompositionally coherent approaches, for subcompositions restricted to containing Si. We thus evaluate the performance of the subcompositionally coherent proposed method $\boldsymbol{\sigma}_0$ approach, by comparing it with classical alr with imputed zeros. For the imputation method, we chose the widely recommended multiplicative replacement method by Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn, 2003, which replaces zeros by a small value and proportionally adjusts the non-zero components so that the relative proportions among the non-zero parts remain unchanged. With default settings, the multRepl command of the package zCompositions removes parts with more than 80% of zeros, This is done to avoid imputing variables that are too sparse, which could lead to unstable results. Here, it removes the part Ba. We therefore included in the comparison both the alr-transformed imputed subcomposition, with Ba removed, and the full alr-transformed imputed composition.

6.2. **Experimental setup.** We thus compare three distinct pipelines:

(1) $\boldsymbol{\sigma}_0$-ratios (6) of the full added oxides (Na, Mg, Al, K, Ca, Ba, Fe) w.r.t. Si;
(2) alr-ratios on the imputed subcomposition (Na, Mg, Al, K, Ca, Fe) w.r.t. Si, with the zero-rich Ba removed;
(3) alr-ratios on the imputed full composition (Na, Mg, Al, K, Ca, Ba, Fe) w.r.t. Si.

In a first experiment, we split the data into a training set (80%) and a test set (20%), and fitted a classical linear regression model on the training set for the prediction of the refractive index on the test set. The process was repeated 100 times. The results are displayed in Table 2.

To allow for a more robust comparison, we conducted a second experiment by using 10-fold cross-splitting: the data is partitioned into 10-non overlapping folds, the models are trained on all observations, except those in the current fold i, and are evaluated by computing the MSE on the observations in fold i. The average MSE on all folds is then computed and the variability of the MSE estimates can be assessed by examining its standard deviation. This is a more efficient method for comparing predictive models: each observation is tested exactly once, giving a direct

[8]available at "https://archive.ics.uci.edu/ml/machine-learning-databases/glass/glass.data". Also available in the R packages mlbench and MASS.

Method	Mean MSE ($\times 10^{-6}$)	SD MSE ($\times 10^{-6}$)	Median MSE ($\times 10^{-6}$)
(1) $\boldsymbol{\sigma}_0$ model	1.129	0.444	1.068
(2) alr+imputation sub	2.766	1.163	2.624
(3) alr+imputation full	1.967	0.706	1.937

TABLE 2. Monte Carlo MSE comparison results between coherent $\boldsymbol{\sigma}_0$ (pipeline (1)) and alr+imputation methods (pipelines (2) and (3)). (100 random splits)

estimate of how the model performs on unseen data from the same distribution, with minimal redundancy. The results are displayed in Table 6.3.

Method	Mean MSE ($\times 10^{-6}$)	SD ($\times 10^{-6}$)
(1) $\boldsymbol{\sigma}_0$ model	1.077	0.594
(2) alr+imputation sub	2.474	1.605
(3) alr+imputation full	1.851	0.855
(4) Simple linear model	1.077	0.597

TABLE 3. Cross-validated MSE comparison (10 folds) between coherent $\boldsymbol{\sigma}_0$ (pipeline (1)) and alr+imputation methods (pipelines (2) and (3)). $\boldsymbol{\sigma}_0$ achieves the same performance as industry's standard approach based on a linear regression model on the components (4), while providing subcompositional coherence.

6.3. **Results and comparison of subcompositionally coherent methods.** Both experimental comparison setups give consistent results: the $\boldsymbol{\sigma}_0$ method yields significantly lower predictive MSE on the test sets than the alr methods, with a mean MSE of $\approx 1.1 \times 10^{-6}$ for our proposed method (1) in both experimental setups, versus 1.96×10^{-6} and 1.85×10^{-6} in the Monte Carlo and CV setup, respectively, for the alr on the full composition (3). Using the alr imputed ratios on the subcomposition excluding Ba, that is, pipeline (2), which is the setup given when using the default settings of the imputation method, yields far worse results ($\approx 2.8\times 10^{-6}$ and $\approx 2.5\times 10^{-6}$ in the Monte Carlo and CV setup, respectively). The standard deviation of the MSE is also smaller for $\boldsymbol{\sigma_0}$ ($\approx 0.4.\times 10^{-6}$ and $\approx 0.6\times 10^{-6}$ in the Monte Carlo and CV setup), which means that σ_0 is also better in terms of stability. In particular, pipeline (2) appears highly unstable. A paired t-test on the cross-validated MSE setup confirms the significance of the difference, yielding the following p-values: $p = 0.0049$ for comparison of (1) with (2), $p = 0.0035$ for comparison of (1) with (3).

It is interesting to compare these statistical results with domain knowledge from the field of Glass Science (see e.g. Scholze, 1991 Chapter 3.4.1, Bach and Neuroth, 1998 Chapter 2.2, and also Lu, Vienna, and Du, 2024). The current standard practice in the industry is to compute the refraction index based on simple linear models on the composition, after pioneering works by Appen, 1949, Huggins, 1940, Huggins and Sun, 1943, as it is argued from physical principles that the refractive index is additive in the individual oxide proportions (Scholze, 1991 p. 223). As

this methodology is based on the components (including the formative oxide Si) and not on their ratios, it is noteworthy that the CoDa's school principle of subcompositional coherence is thus not taken into account in the industry standard practice. Consequently, we performed a simple linear regression on the full composition, see line (4) in Table 6.3. It is remarkable that our proposed method, based on $\boldsymbol{\sigma}_0$ ratios, achieve the same MSE performance 1.077×10^{-6} on this dataset as the recommended methodology of the industry, (even with a slightly better SD), while adding the principle of subcompositional coherence, and clearly outperforming methods based on (imputed) log-ratios.

6.4. **Coefficients comparison.** Regarding the coefficients in the models, Table 4 compares the alr_{Si} coefficients for the subcomposition (2) and the full composition (3), whereas Table 4 compares the $\boldsymbol{\sigma}_{Si}$ coefficients of pipeline (1) versus the alr_{Si} coefficients for the full composition (3).

Predictor	Mean	SD
(Intercept)	1.57884	0.00411
$\text{alr}_{Si}(Ca)$	0.01975	0.00110
$\text{alr}_{Si}(Na)$	0.01046	0.00142
$\text{alr}_{Si}(Mg)$	0.00052	0.00012
$\text{alr}_{Si}(K)$	0.00027	0.00007
$\text{alr}_{Si}(Fe)$	0.00010	0.00009
$\text{alr}_{Si}(Al)$	-0.00073	0.00024

Predictor	Mean	SD
(Intercept)	1.58502	0.00254
$\text{alr}_{Si}(Ca)$	0.02212	0.00075
$\text{alr}_{Si}(Na)$	0.00809	0.00102
$\text{alr}_{Si}(Ba)$	0.00100	0.00010
$\text{alr}_{Si}(Mg)$	0.00096	0.00008
$\text{alr}_{Si}(K)$	0.00049	0.00006
$\text{alr}_{Si}(Fe)$	-0.00004	0.00004
$\text{alr}_{Si}(Al)$	-0.00166	0.00018

TABLE 4. Comparison of coefficient estimates (mean and standard deviation) for the alr sub-model (2) (left) and the full alr model (3) (right), over 100 random splits.

Predictor	Mean	SD
(Intercept)	1.47779	0.00124
$\boldsymbol{\sigma}_{Si}(Ca)$	0.23398	0.00438
$\boldsymbol{\sigma}_{Si}(Ba)$	0.18049	0.01071
$\boldsymbol{\sigma}_{Si}(Mg)$	0.09034	0.00437
$\boldsymbol{\sigma}_{Si}(Na)$	0.07614	0.00450
$\boldsymbol{\sigma}_{Si}(K)$	0.06296	0.00751
$\boldsymbol{\sigma}_{Si}(Fe)$	0.00511	0.03665
$\boldsymbol{\sigma}_{Si}(Al)$	-0.04872	0.00990

Predictor	Mean	SD
(Intercept)	1.58502	0.00254
$\text{alr}_{Si}(Ca)$	0.02212	0.00075
$\text{alr}_{Si}(Na)$	0.00809	0.00102
$\text{alr}_{Si}(Ba)$	0.00100	0.00010
$\text{alr}_{Si}(Mg)$	0.00096	0.00008
$\text{alr}_{Si}(K)$	0.00049	0.00006
$\text{alr}_{Si}(Fe)$	-0.00004	0.00004
$\text{alr}_{Si}(Al)$	-0.00166	0.00018

TABLE 5. Comparison of coefficient estimates (mean and standard deviation) for the $\boldsymbol{\sigma}_0$ model (1) (left) and the alr model (3) (right) over 100 random splits.

Although the coefficients are not directly comparable across methods, as they are transformed by different functions, some interesting points are to be noted. For the additive log-ratio methods (Table 4), excluding Ba on the ground of having too

many zeros, as is often the methodology suggested in the literature and is the output generated by the multiplicative replacement input method using default settings is problematic: in the full composition, the log-ratio $\text{alr}_{Si}(Ba)$ appears in third most influential oxide ratio in impacting the Refractive Index (while the remaining order for the other oxides remain the same). This suggests that oxides in small proportions may also be important for predicting the dependent variable. Interestingly, the sign of $\text{alr}_{Si}(Fe)$ changes sign between (2) and (3), which evidences instability. In addition, the standard deviation of the alr_{Si} coefficients in the subcomposition excluding Ba (2) are always higher than the corresponding ones on the full composition (3), which indicates more unstable estimates in the sub-composition (2).

Regarding comparison of the $\boldsymbol{\sigma}_{Si}$ coefficients of pipeline (1) versus the alr_{Si} coefficients for the full composition (3) (Table 5), it is noteworthy that the proposed approach (1) ranks the contribution of Ba in second place, with $\boldsymbol{\sigma}_{Si}(Ba) = 0.18$ close to the first contributing factor $\boldsymbol{\sigma}_{Si}(Ca) = 0.23$, whereas the log-ratio approaches degrades its importance to third place, with $\text{alr}_{Si}(Ba) = 0.0010$ considerably lower than the first two contributing factors $\text{alr}_{Si}(Ca) = 0.022$ and $\text{alr}_{Si}(Na) = 0.0080$. Also, (Mg) is relegated to the fourth place, with $\text{alr}_{Si}(Mg) = 0.000096$, and in particular is ranked after (Na).

Leveraging domain knowledge, it is known that Barium oxide (BaO) has historically played a crucial role in the manufacturing of optical glass, due to its ability, as a heavy alkaline earth metals with a very large ionic radius and high polarizability, to significantly increase the refractive index of glass while keeping optical dispersion relatively low (a property discovered in 1884 by Otto Schott, per Encyclopedia Britannica). Scholze, 1991 Table 29 column 3 p. 224 lists Barium's refractivity coefficient (1.880) as significantly higher than Sodium (Na)'s (1.590) or Magnesium's (1.610), in Appen's model (see also Bach and Neuroth, 1998 Table 2.5 p. 74). This domain-knowledge again suggests that the proposed model (1) based on $\boldsymbol{\sigma}_0$ ratios better captures the fundamental properties of the oxides involved than those based on log-ratios, in particular giving the order Ba > Mg > Na, consistent with the literature.

6.5. **Comparison with global representations.** Although our primary focus is in comparing subcompositionally coherent methods, we include in Table 6.5, for comparison sake, the results obtained by performing an OLS regression based on the proposed global representations $\boldsymbol{\sigma}_*$ of Equation (13) and π_0^+ of (15). The MSE were calculated on the same folds as those of Table 6.3. Here, the performance is still very close to the $\boldsymbol{\sigma}_0$ method (1) and the industry's method (4) of Table 6.3, and also clearly outperform the log-ratio with imputation methods (2) and (3). However, and in contrast to method (1) based on $\boldsymbol{\sigma}_0$, methods based on $\boldsymbol{\sigma}_*$ or π_0^+ no longer satisfy the subcompositional coherence properties of Proposition 5.2.

Method	Mean MSE ($\times 10^{-6}$)	SD ($\times 10^{-6}$)
(5) $\boldsymbol{\sigma}_*$	1.080	0.608
(6) π_0^+	1.0784	0.599

TABLE 6. Cross-validated MSE comparison (10 folds) between global representations $\boldsymbol{\sigma}_*$ of Equation (13) and the variant π_0^+ of (15).

6.6. **Model Diagnostics.** At last, we compare and contrast in Figures 7 and 8 the diagnostics plots obtained using R's package performance for the simple linear model (4) of industry's practice and the proposed linear model based on $\boldsymbol{\sigma}_0$ ratios (1). The models are recalculated on the whole sample.

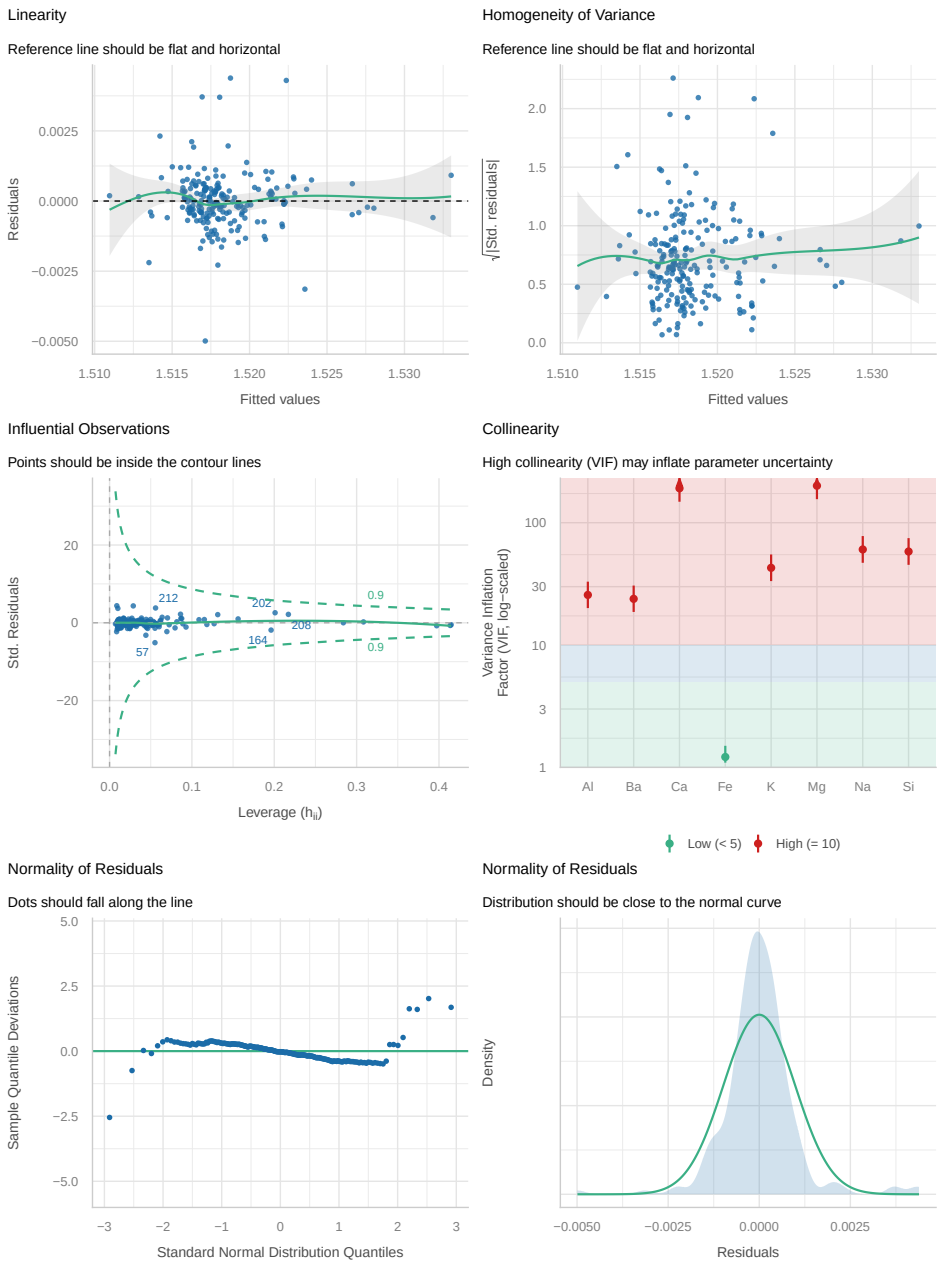

FIGURE 7. Diagnostic plots for the simple linear model (4)

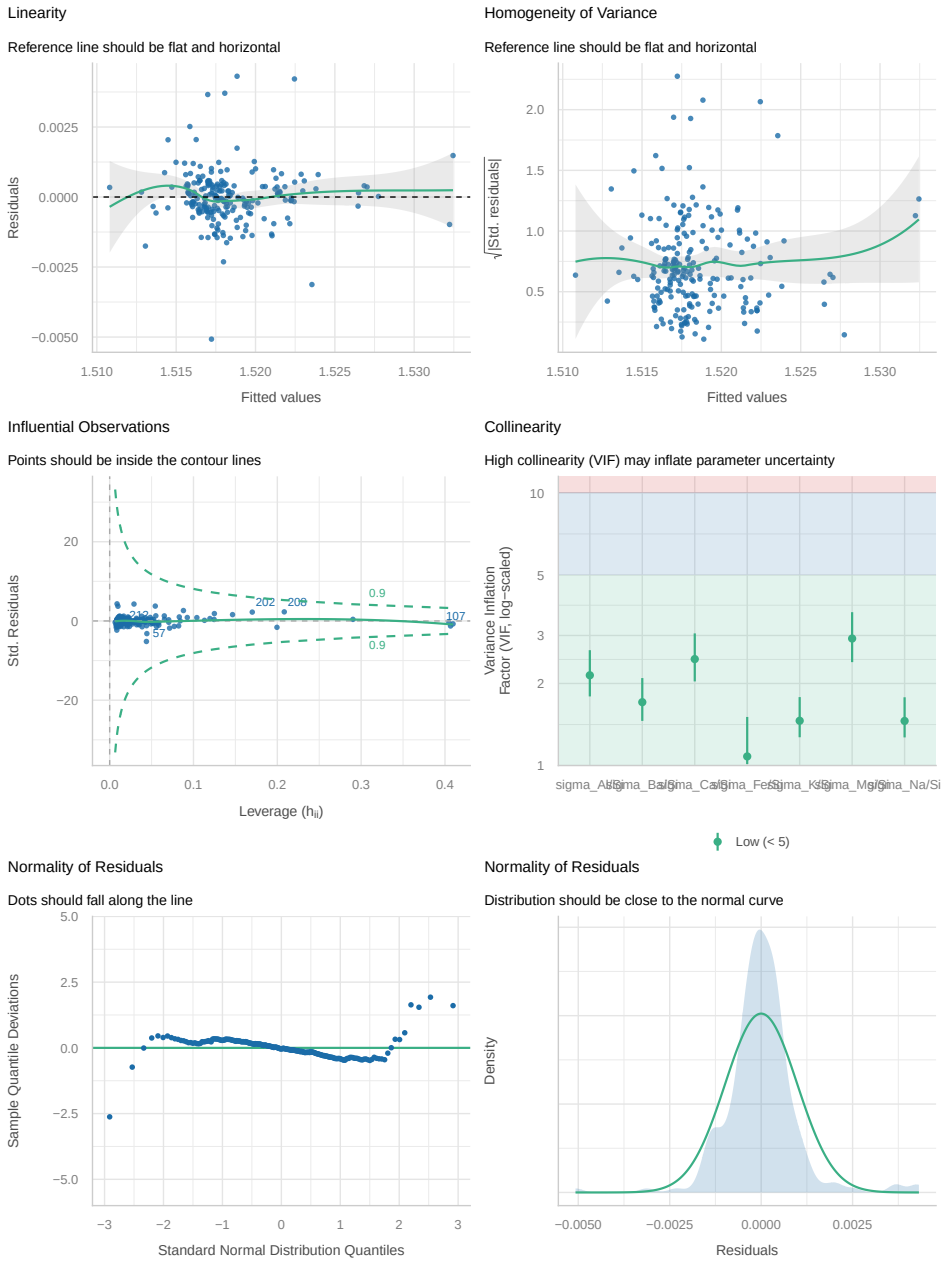

FIGURE 8. Diagnostic plots for the $\boldsymbol{\sigma}_0$ model (1)

Both models show similar and satisfactory diagnostic plots regarding linearity, homoscedasticity, influential observations and normality of residuals, suggesting an adequate fit to the data. A notable difference is in the plot of the Variance Inflation Factor (VIF) (middle right panel): the linear model (4) exhibits severe

multi-collinearity which harms inference and makes it difficult to estimate and interpret the unique effect of each predictor. This is to be expected since the predictors in the plain linear model are the components, which sum to one. This induces the well-known issue in CoDa analysis of spurious correlation. To the contrary, the VIF in model (4) based on $\boldsymbol{\sigma}_0$-ratios are all below 3, which is generally considered as acceptable. This vindicates the use of ratios for CoDa analysis and suggests that the proposed method (1) based on $\boldsymbol{\sigma}_0$-ratios is advantageous w.r.t. to industry's standard method (4).

7. Conclusion

In the introduction we asked, if there exist CoDa representations,

i) able to deal with zeroes,
ii) which are easily interpretable in terms of ratios of components,
iii) which (at least partially) maintain subcompositional coherence,
iv) and which are based on an easy framework of simply transforming the data.

Investigating the ratio-based homeomorphism σ of (2), in order to deal with ratios with zeros as extended numbers, led us to introduce both i) the local representations $\boldsymbol{\sigma}_0$ of (6), and $\boldsymbol{\sigma}_{0\dots k}$ of (10) and ii) the global representations $\boldsymbol{\sigma}_*$ of (13) and π_0^+ of (15). We discovered that no single transformation satisfy all desirable properties: ability to handle zeros globally, zero-coherence, and subcompositional coherence. In particular, we evidenced a trade-off between the breadth of handling zeros and the subcompositional coherence property, with classical log-ratios being fully coherent but unable to intrinsically handle any zeros on one side, and global representations ii), like our $\boldsymbol{\sigma}^*$ or π_0^+, which are subcompositionally incoherent but able to handle zeros in any components. A particularly interesting compromise is obtained by the local representations i) $\boldsymbol{\sigma}_0$ and $\boldsymbol{\sigma}_{0\dots k}$, which allow for a quasi-global treatment of zeros and (partial) coherence. All coordinate representations have an interesting interpretation as ratios of components or "excesses" of the components.

A numerical illustration on predicting the Refractive Index of glass data showed that statistical analysis based on such local representation $\boldsymbol{\sigma}_0$ of the composition allowed to match the performance of the industry's standard linear methods, while gaining subcompositional coherence, and outperforming log-ratios with imputation methods. One thus gain a family of representations of CoDa with zeros as points embedded in subsets of Euclidean space, on which classical multivariate techniques can be performed, without the need for instability-inducing imputation methods.

Let us close this conclusion with added discussion which provide additional insight and may stimulate further research.

8. Supplementary discussion

8.1. **Revisiting the principle of scale invariance for CoDa with zeros.** We reminded in Section 2, that (Aitchison, 1986; Aitchison, 1992)'s principle of scale invariance lead to the analysis of properties expressed as scale invariant functions. More generally, let us recall that a function $f : V \to W$ between two two vector spaces V, W is homogeneous (resp. positive homogeneous) of degree k if for all $\lambda \in \mathbb{R}$ (resp. $\lambda > 0$), $\mathbf{x} \in V$,

$$f(\lambda \mathbf{x}) = \lambda^k f(\mathbf{x}), \tag{17}$$

for some $k \in \mathbb{R}$ (resp. $k \geq 0$). In particular, a homogeneous function of degree 0 is scale-invariant. Differentiable positive homogeneous functions are characterized by Euler's formula (1755), see e.g. (Aczél, 1966):

Theorem 8.1 (Euler's homogenous function theorem)**.** *If f is a function of $d+1$ variables, and continuously differentiable in some open subset of $\mathbb{R}^{d+1}$, then f is positive homogeneous of degree k if and only if it satisfies Euler's partial differential equation,*

$$\sum_{i=0}^{d+1} x_i . \frac{\partial f}{\partial x_i}(x_0, \ldots, x_d) = k f(x_0, \ldots, x_d).$$

For example, it is easy to check that a simple ratio

$$f(\mathbf{x}) := \frac{x_i}{x_0}$$

satisfy Euler's equation:

$$x_0 \frac{\partial f}{\partial x_0} + x_i \frac{\partial f}{\partial x_i} = x_0 \frac{-x_i}{x_0^2} + x_i \frac{1}{x_0} = 0$$

for $x_0 \neq 0$. From Euler's theorem, one can deduce the general form of a smooth positively homogeneous function for $x_0 \neq 0$ as (see Aczél, 1966 p. 325 in the bivariate case)

$$f(x_0, \ldots, x_d) = x_0^k \phi(x_1/x_0, \ldots, x_d/x_0), \quad x_0 \neq 0, \tag{18}$$

for some smooth function ϕ. Thus, a differentiable scale invariant function, writes, for $x_0 \neq 0$, as a function of the d-ratios $x_1/x_0, \ldots, x_d/x_0$, hereby confirming, by Euler's equation, the mandate, spelled by Aitchison from invariant theory, that CoDa analysis must be based on (and only on) ratios of components.

However, the important point to note here is that such a representation (18) only works *locally*, for points with $x_0 \neq 0$. An essential distinction and complications occur when there are zeros: if one insists on using ratios of components, one must abandon the idea of using a single ratio representation[9], as several functions are needed to cover the whole CoDa space in order to account for its stratification along the pattern of zero and non-zero components. In particular, a ratio $f = P/Q$ of homogeneous functions P, Q of same degree (e.g. homogeneous polynomials of same degree) gives a scale-invariant function, which is *only properly defined on the non-zero set of Q*.

The foregoing discussion sheds light and puts in perspective the results obtained: $\boldsymbol{\sigma}_0$ is a local representation for CoDa without zeros in the first component x_0 (Lemma 3.1), which by virtue of its ratio form involving only parts 0 and i, is subcompositionally coherent for subcompositions containing part 0 (Proposition 5.2). To the contrary, the global representations $\boldsymbol{\sigma}^*$ and π_0^+ of Section 4.2 are of the form P/Q with Q never vanishing (Lemmas 4.1 and 4.2). Globality is obtained by having as denominator Q the "excess" $1 + x_0 = \sum_{j=0}^{d} x_j + x_0$ involving all components. By this feature, it is never vanishing but can not be subcompositionally coherent.

In addition, this discussion shows that one can somehow enlarge the commonly accepted definition of what constitutes "ratios of components". Indeed, realizing that for a CoDa $\mathbf{x} \in \Delta_+^d$ represented on the simplex, the constant 1 is the sum of the components, and thus that ratios of linear combinations of components can

[9]This is the starting point of the idea of a manifold.

also include 1 and still remain ratios of linear combination of components. This means that *affine* combinations of simplex components, $k+\sum_{j=0}^{d}\alpha_j x_j$, with integer $k \in \mathbb{N}$, are in fact *linear* combination of CoDa simplex components.

How trivial this observation may be, it is somehow unintuitive in regards of the ratio-scale (Stevens, 1946) attributed to CoDa. Indeed, for a measurement on a ratio scale (e.g. length), which has a real zero point in the sense that zero measurement represents the absence of the property being measured, one can make ratio statements about the property (e.g. saying that a person is twice as high as another). Multiplication by a constant does not destroy its ratio character (e.g. the same person remaining twice as high, whether height being measured in cm or in inches), but addition of a constant does. This may explain why our proposed solutions $\boldsymbol{\sigma}_*$ and π_0^+ to deal with CoDa with zeros (add one to the denominator and/or the numerator ratios) have remained elusive in the CoDa literature for so many years, in spite of their simplicity and of being widely known (the shifting or "starting" trick of Tukey, 1977) in classical multivariate analysis). Note also that we have argued elsewhere (Faugeras, 2026, Faugeras, 2025) that such ratio scale breaks down for CoDa with zeros and that CoDa have a more complex projective/affine structure.

8.2. **On σ transform of log-ratios for CoDa with zeros.** For simplicity and maintaining the unity and focus of the paper, we have dealt principally with ways to obtain surrogate *ratios* for CoDa with zeros, via the transform σ and the deduced representations. A more principled approach for obtaining surrogates of *log-ratios*, in particular Aitchison's log-ratios (alr,clr,ilr) transforms, for CoDa with zeros will be be dealt with in a subsequent paper, using a different generalization of the homeomorphism σ. Nonetheless, let us quickly mention how the σ homeomorphism of Section 2 can also be applied to several log-ratios. This gives expedient and pragmatic solutions for extending log-ratios representations to CoDa with zeros.

Indeed, when applied σ of (2) is applied this time to a log-ratio, the extended number $\ln(x_i/x_j) \in [-\infty, \infty]$ turns into

$$\sigma\left(\ln\left(\frac{x_i}{x_j}\right)\right) = \frac{\ln(x_i/x_j)}{1+|\ln(x_i/x_j)|} \in [-1,1],$$

which is also a bounded, scale invariant quantity. Reasoning as in Section 3.1, but for log ratios

$$\mathbf{ln}\,\pi_0([\mathbf{x}]_+) := \left(\ln\left(\frac{x_1}{x_0}\right), \ldots, \ln\left(\frac{x_d}{x_0}\right)\right) \in [-\infty,\infty]^d,$$

one can apply to them the homeomorphism σ, yielding a mapping

$$\begin{aligned}\mathbf{L}_0 : \mathbb{P}_+^d &\to L_0^d := \mathbf{L}_0(\mathbb{P}_+^d) \subset [-1,1]^d\\ [\mathbf{x}]_+ &\mapsto \mathbf{L}_0([\mathbf{x}]_+) := (y_1,\ldots,y_d)\end{aligned}$$

with

$$y_i = \sigma \circ \ln(x_i/x_0) = \frac{\ln(x_i/x_0)}{1+|\ln(x_i/x_0)|}, \quad i = 1,\ldots,d.$$

This corresponds to applying the σ homeomorphism to Aitchison's alr transform.

One has a very similar discussion as in the previous Section 3.1: although defined on the whole CoDa space, $\mathbf{L}_0$ sends CoDa points with $x_0 = 0$, $x_i \neq 0$, $i = 1,\ldots,d$ to the same $\mathbf{1} \in \mathbb{R}^d$ (while $y_i = 0$ if $x_0 = x_i = 0$ with the convention $0/0 := 1$) and is thus not injective on the whole CoDa space, but only on $\mathbb{P}_+^d \setminus \mathbb{U}_0$. This

gives a local ratios of log-ratios representation for CoDa with zeroes outside the x_0 component, which is interpreted as an extension of the alr transform. Variant to the log of the ratios $\pi_{0\dots k}$ of (9) in Section 3.2 could also be envisioned.

Let us simply state (without proof) the analogue of Lemma 3.1:

Lemma 8.2. $\mathbf{L}_0$ *realizes an embedding from* $\mathbb{P}_+^d \setminus \mathbb{U}_0^d$ *into the unit cube* $[-1,1]^d$, *viz.* $\mathbf{L}_0 : \mathbb{P}_+^d \setminus \mathbb{U}_0^d \hookrightarrow [-1,1]^d$, *and is bijective from* $\mathbb{P}_+^d \setminus \mathbb{U}_0^d$ *onto* $[-1,1)^d$, *with inverse transformation*

$$(\mathbf{L}_0)^{-1} : [-1,1)^d \to \mathbb{P}_+^d \setminus \mathbb{U}_0^d$$
$$(y_1, \dots y_d) \mapsto [\mathbf{x}]_+,$$

where $\mathbf{x} \in \Delta_+^d$ *is given by*

$$x_i = \frac{e^{t_i}}{1 + \sum_{j=1}^d e^{t_j}}, \quad i = 1, \dots d$$
$$x_0 = \frac{1}{1 + \sum_{j=1}^d e^{t_j}}, \quad \text{with } t_i = \frac{y_i}{1 - |y_i|}$$

Figure 9 illustrates the construction for $d = 2$. The isoline $x_0 = 0$ is sent to the single (red) point $(1, 1)$, except for the vertices with $x_1 = 0$ and $x_2 = 0$ which are sent, resp., to $(0, 1)$ and $(1, 0)$ (red dots). For $x_0 \neq 0$, points with $x_1 = 0$, resp $x_2 = 0$, are sent to the lower, resp. left, side of the cube $[-1, 1)^2$, as in Figures 2 and 3.

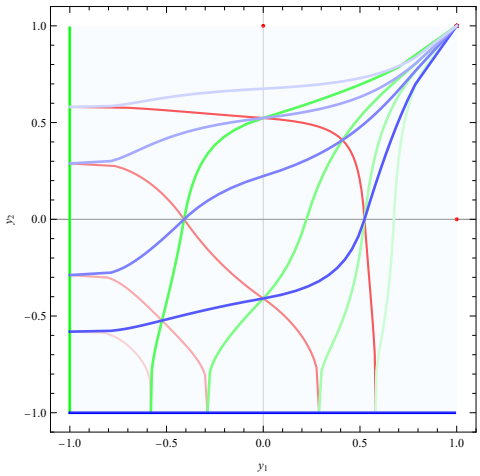

FIGURE 9. Region plot of $L_0^2 = [-1, 1)^2$ (lightblue square), with images by $\mathbf{L}_0$ of the iso-lines for x_0 (red), x_1 (green), x_2 (blue) of the simplex of Figure 2, for $d = 2$ (3 components).

Regarding the transformations of Section 4, the σ homeomorphism can be applied to the ln coordinates of the simplex, yielding

$$\mathbf{L}_* : \mathbb{P}_+^d \to L_*^d := \mathbf{L}_*(\mathbb{P}_+^d) \subset [-1,1]^d$$
$$[\mathbf{x}]_+ \mapsto \mathbf{L}_*([\mathbf{x}]_+) := \sigma \circ \mathbf{ln} \circ \mathcal{C}(\mathbf{x}) = \left(\frac{\ln x_1}{1 + |\ln x_1|}, \dots, \frac{\ln x_d}{1 + |\ln x_d|} \right).$$

For the global variant representation π_0^+ of (15), it appears more useful to apply the ln function to the always positive ratios $(1+x_i)/(1+x_0)$: if one defines $\mathbf{L}_0^+ := \mathbf{ln} \circ \pi_0^+$, viz.

$$\begin{aligned} \mathbf{L}_0^+ : \mathbb{P}_+^d &\to \mathbb{R}^d \\ [\mathbf{x}]_+ &\mapsto \left(\ln \frac{1+x_1}{1+x_0}, \ldots, \ln \frac{1+x_d}{1+x_0} \right), \end{aligned}$$

one obtains a symmetric, zero-coherent global representation. This corrects the lack of zero-coherence of π_0^+. The latter is also interpreted as a global extension (this time) of the alr_0 transform for CoDa with zeros. However, and compared to $\boldsymbol{\sigma}_0$, neither $\mathbf{L}_0$ nor $\mathbf{L}_0^+$ are subcompositionally coherent, as they involve all coordinates through the constant $1 = \sum_{j=0}^d x_j$.

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.

Appendix A. Proofs

Proof of Lemma 3.1. Let $[\mathbf{x}]_+, [\mathbf{y}]_+ \in \mathbb{P}_+^d \setminus \mathbb{U}_0^d$.

$$\begin{aligned} \boldsymbol{\sigma}_0([\mathbf{x}]_+) = \boldsymbol{\sigma}_0([\mathbf{y}]_+) &\iff \frac{x_i}{x_0+x_i} = \frac{y_i}{y_0+y_i}, \quad i = 1, \ldots, d \\ &\iff x_i y_0 = y_i x_0, \quad i = 1, \ldots, d \\ &\iff \begin{cases} x_i = y_i = 0 \text{ if } x_i = 0 \text{ or } y_i = 0 \text{ for some } i, \\ \text{or} \\ x_i, y_i \neq 0 \text{ and } \frac{x_i}{y_i} = \frac{x_0}{y_0} \end{cases} \\ &\iff [\mathbf{x}]_+ = [\mathbf{y}]_+, \end{aligned}$$

since $x_0, y_0 \neq 0$. This proves injectivity. $x_0 > 0$ entails $0 \leq x_i/(x_0+x_i) < 1$, thus $\boldsymbol{\sigma}_0(\mathbb{P}_+^d \setminus \mathbb{U}_0^d) \subset [0,1)^d$. Conversely, given $\mathbf{z} \in [0,1)^d$, $0 \leq z_i < 1$, for all $i = 1, \ldots, d$, entails that $\mathbf{x}$, defined by (7), is well-defined as a real vector, belong to Δ_+^d and satisfy $\boldsymbol{\sigma}_0([\mathbf{x}]_+) = \mathbf{z}$. □

Remark 5. *Note that $\boldsymbol{\sigma}_0^{-1}$ (equation (7)) extends, with the convention $0/0 := 1$, to the whole of $[0,1]^d$, sending any $\mathbf{z}$ with k components equal to 1 to $\mathbf{x} \in \Delta_+^d$ with $x_0 = 0$, $x_i = 1/k$ for the parts i with $z_i = 1$, and $x_i = 0$ for the parts i with $z_i < 1$. In particular, $\boldsymbol{\sigma}_0^{-1}(\mathbf{1}) = (0, 1/d, \ldots, 1/d)$.*

Proof of Lemma 4.1. Similar to that of Lemma 3.1. □

Proof of Lemma 4.2. Bijectivity is established by the formula of the inverse transformation, which follows by simple algebraic manipulation and is well defined, since $\sum_{j=1}^d z_i > 0$. □

Proof of Proposition 5.1. The proof follows from the corresponding zero-coherence of the ℓ_p norm of Euclidean vectors, viz. $\|\mathbf{z} : \mathbf{0}\|_p = \|\mathbf{z}\|_p$ for $\mathbf{z} \in \mathbb{R}^d$ and $\mathbf{0} \in \mathbb{R}^k$, and the fact that $\sigma(0) = 0$. □

Proof of Proposition 5.2. i) From (6), it is clear that each component z_i of $\boldsymbol{\sigma}_0([\mathbf{x}]_+)$ only depends on components x_0, x_i of $\mathbf{x}$. Thus, by scale invariance of $\boldsymbol{\sigma}_0$, one obtains that the $\boldsymbol{\sigma}_0$-representation of the subcomposition $[\mathcal{S}^{\Delta}_{\{0\}\cup I}(\mathbf{x})]_+$ is the same as the selection of the components in I of the $\boldsymbol{\sigma}_0$-representation of $[\mathbf{x}]_+$, viz.

$$\boldsymbol{\sigma}_0(\mathcal{S}_{\{0\}\cup I}([\mathbf{x}]_+) = \mathcal{S}_I(\boldsymbol{\sigma}_0([\mathbf{x}]_+))$$

ii) Since $d_p^{\boldsymbol{\sigma}_0}$ is the ℓ_p distance in the $\boldsymbol{\sigma}_0$-representation, and the ℓ_p distance between subvectors is lower than the distance between the original vectors, i.e. $\|\mathcal{S}_I(\mathbf{u}) - \mathcal{S}_I(\mathbf{v})\|_p \leq \|\mathbf{u} - \mathbf{v}\|_p$, for any vectors $\mathbf{u}, \mathbf{v}$, the result follows from i).

□

References

Aczél, J. (1966). *Lectures on Functional Equations and Their Applications*. Mathematics in Science and Engineering, Vol. 19. Translated by Scripta Technica, Inc. Supplemented by the author. Edited by Hansjorg Oser. Academic Press, New York-London, pp. xx+510.

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

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.

Aitchison, John (1992). "On criteria for measures of compositional difference". In: *Math. Geol.* 24(4), pp. 365–379. ISSN: 0882-8121. DOI: 10.1007/BF00891269.

Appen, AA (1949). "Calculation of optical properties, density and thermal expansion of silicate glasses on the base of their composition". In: *Dokl. AN SSSR*. Vol. 69. 6, pp. 841–4.

Bach, Hans and Norbert Neuroth, eds. (1998). *The Properties of Optical Glass*. Schott Series on Glass and Glass Ceramics. Springer: Berlin. DOI: 10.1007/978-3-642-57769-7.

Bourbaki, N. (1960). *Éléments de mathématique. Première partie. (Fascicule III.) Livre III; Topologie générale. Chap. 3: Groupes topologiques. Chap. 4: Nombres réels*. Actualités Scientifiques et Industrielles [Current Scientific and Industrial Topics], No. 1143. Troisième édition revue et augmentée. Hermann, Paris, 236 pp. (1 insert).

Faugeras, Olivier P. (Dec. 2023). "An invitation to intrinsic compositional data analysis using projective geometry and Hilbert's metric". 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. (June 2025). "Log-Free Distance and Covariance Matrix for Compositional Data II: The Projective/Exterior Product Approach". In: *Austrian Journal of Statistics* 54, pp. 122–160. URL: https://ajs.or.at/index.php/ajs/article/view/2105.

Faugeras, Olivier P. (Mar. 2026). "Log-Free Divergence and Covariance matrix for Compositional Data I: The Affine/Barycentric Approach". In: *Austrian Journal of Statistics*. Accepted. URL: https://hal.science/hal-04670796v2.

Gajer, Pawel and Jacques Ravel (2025). *A New Approach to Compositional Data Analysis using L^∞-normalization with Applications to Vaginal Microbiome*. arXiv: 2503.21543 [stat.CO]. URL: https://arxiv.org/abs/2503.21543.

Greenacre, Michael (2018). *Compositional data analysis in practice*. CRC press: New York.

Greenacre, Michael (2021). "Compositional data analysis". In: *Annu. Rev. Stat. Appl.* 8, pp. 271–299. ISSN: 2326-8298. DOI: 10.1146/annurev-statistics-042720-124436.

Huggins, Maurice L (1940). "The refractive index of silicate glasses as a function of composition". In: *Journal of the Optical Society of America* 30(10), pp. 495–504.

Huggins, Maurice L. and Kuan -Han Sun (1943). "Calculation of Density and Optical Constants of a Glass from its Composition in Weight Percentage". In: *Journal of the American Ceramic Society* 26(1), pp. 4–11. URL: https://ceramics.onlinelibrary.wiley.com/doi/abs/10.1111/j.1151-2916.1943.tb15176.x.

Lu, Xiaonan, John D Vienna, and Jincheng Du (2024). "Glass formulation and composition optimization with property models: A review". In: *Journal of the American Ceramic Society* 107(3), pp. 1603–1624.

Martín-Fernández, Josep A, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn (2003). "Dealing with zeros and missing values in compositional data sets using nonparametric imputation". In: *Mathematical Geology* 35(3), pp. 253–278.

Mosteller, Frederick and John W Tukey (1977). *Data analysis and regression. A second course in statistics*. Addison-Wesley series in behavioral science: quantitative methods.

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.

Pearson, Karl (1897). "Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs". In: *Proceedings of the royal society of london* 60(359-367), pp. 489–498.

Richter-Gebert, Jürgen (2011). *Perspectives on projective geometry*. A guided tour through real and complex geometry. Springer, Heidelberg, pp. xxii+571. ISBN: 978-3-642-17285-4. DOI: 10.1007/978-3-642-17286-1.

Scealy, J. L. and A. H. Welsh (2014). "Colours and cocktails: compositional data analysis 2013 Lancaster lecture". In: *Aust. N. Z. J. Stat.* 56(2), pp. 145–169. ISSN: 1369-1473. DOI: 10.1111/anzs.12073.

Scholze, Horst (1991). *Glass: Nature, Structure, and Properties*. Springer-Verlag: New York. DOI: 10.1007/978-1-4612-3032-8.

Stevens, S. S. (1946). "On the Theory of Scales of Measurement". In: *Science* 103(2684), pp. 677–680. ISSN: 00368075, 10959203. URL: http://www.jstor.org/stable/1671815.

Tsagris, Michail (2015). "Regression analysis with compositional data containing zero values". In: *Chil. J. Stat.* 6(2), pp. 47–57. ISSN: 0718-7912.

Tsagris, Michail, Simon Preston, and Andrew TA Wood (2016). "Improved classification for compositional data using the α-transformation". In: *Journal of classification* 33, pp. 243–261.

Tsagris, Michail T, Simon Preston, and Andrew TA Wood (2011). "A data-based power transformation for compositional data". In: *arXiv preprint arXiv:1106.1451*.

Tukey, John Wilder (1977). *Exploratory data analysis*. Vol. 2. Springer.

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

URL: https://www.tse-fr.eu/fr/people/olivier-faugeras