

Testing for a common subspace in compositional datasets with structural zeros

Francesco Porro^a, Fabio Rapallo^b, Sara Sommariva^{a,*}

^a*Department of Mathematics, University of Genova, via Dodecaneso
35, Genova, 16146, , Italy*

^b*Department of Economics, University of Genova, via Francesco Vivaldi
5, Genova, 16126, , Italy*

Abstract

In real world applications dealing with compositional datasets, it is easy to face the presence of structural zeros. The latter arise when, due to physical limitations, one or more variables are intrinsically zero for a subset of the population under study. The classical Aitchison approach requires all the components of a composition to be strictly positive, since the adaptation of the most widely used statistical techniques to the compositional framework relies on computing the logratios of these components. Therefore, datasets containing structural zeros are usually split in two subsets, the one containing the observations with structural zeros and the one containing all the other data. Then statistical analysis is performed on the two subsets separately, assuming the two datasets are drawn from two different subpopulations. However, this approach may lead to incomplete results when the split into two populations is merely artificial. To overcome this limitation and increase the robustness of such an approach, we introduce a statistical test to check whether the first K principal components of the two datasets generate the same vector space. An approximation of the corresponding null distribution is derived analytically when data are normally distributed on the simplex and through a nonparametric bootstrap approach in the other cases. Results from simulated data demonstrate that the proposed procedure can discriminate scenarios where the subpopulations share a common subspace from those where they are actually distinct. The performance of the proposed method is also tested on an experimental dataset concerning

*sara.sommariva@unige.it

microbiome measurements.

Keywords: Common subspace, Compositional data, Logratio transformations, Nonparametric hypothesis testing, Principal Component Analysis

1. Introduction

The problem we deal with in this paper falls in the framework of dimensionality reduction for compositional data with structural zeros. In recent decades, the number of analyses dealing with compositional data has been largely increased. Compositional data arise in contexts where the relevant information lies in the proportions among the observed variables and not in their values or their sum. Possible fields of application include tourism (Grifoll et al., 2019), finance (Fiori and Porro, 2023), microbiome analysis (Tsilimigras and Fodor, 2016), pattern recognition (Lu et al., 2024), geochemistry and analytical chemistry (Rieser and Filzmoser, 2023; Mert et al., 2015).

At the core of compositional data analysis lies the definition of compositions. A composition is a real-valued vector having all strictly positive components that sum to a fixed value κ . A suitable space to contain all the compositions with D parts is the simplex, defined by:

$$\mathbb{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_D) : x_i > 0, \forall i; \sum_{i=1}^D x_i = \kappa \right\}. \quad (1)$$

In the remaining of the paper we will assume $\kappa = 1$. This is a common choice in compositional data analysis as it allows to identify a composition with a vectors of proportions that sum to 1.

A typical compositional dataset $\mathcal{X}$ is a matrix with n rows and D columns collecting a sample of n observations, each one being a D -part composition:

$$\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)', \text{ where:} \quad (2)$$

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iD}) \in \mathbb{S}^D, \sum_{j=1}^D x_{ij} = 1 \quad i = 1, 2, \dots, n.$$

The standard statistical descriptive measures based on the real Euclidean geometry can lead to erroneous conclusions when applied to compositional

datasets (Pawlowsky-Glahn et al., 2015; Mert et al., 2015; Grifoll et al., 2019). This issue is usually overcome through the so-called *principle of working in coordinates* (Mateu-Figueras et al., 2011; Grifoll et al., 2019): a proper transformation is defined to map each composition into a vector of coordinates belonging to suitable spaces equipped with an Euclidean structure. Then, standard statistical approaches, such as e.g. Principal Component Analysis (PCA) for dimensionality reduction, can be applied to the transformed data.

Many transformations based on the logratios have been proposed, among which the most common ones are the additive logratio (alr), the centered logratio (clr), and the isometric logratio (ilr) (Aitchison, 1982; Pawlowsky-Glahn et al., 2015; Egozcue et al., 2003; Mateu-Figueras et al., 2011). All these three transformations are based on the ratios of logarithms of parts belonging to a composition. From a theoretical perspective, this does not cause issues since, by definition of composition, and also as showed in the simplex formula in Eq. (1), all the parts of a composition must be greater than zero. The point is that, unfortunately, in many real world applications, the present of a null part can occur. In those cases, the aforementioned transformations cannot be applied, so new procedures must be considered.

As detailed in Pawlowsky-Glahn et al. (2015) and Martin-Fernandez et al. (2012), there are distinct kinds of zeros. In this paper we focus on *structural zeros*, which are actual zeros, representing the absence of the phenomenon under analysis. As an example, in a study on monthly expenditure of a set of families, the proportion of expenditure in children school services in families without children is a structural zero. Structural zeros are different from *counting zeros*, which are caused by sampling issues of an unobserved part, or from *rounded zeros*, that are due to a measurement under a certain threshold (Pawlowsky-Glahn et al., 2015; Kim et al., 2024). For a detailed review on the comparison of zero replacement strategies for compositional data see Lubbe et al. (2021).

While, after making some considerations and assumptions, it can sound reasonable to replace the counting and the rounded zeros with a certain (small) value ϵ (Lubbe et al., 2021; Filzmoser et al., 2018), this procedure does not seem acceptable when dealing with structural zeros. Indeed, in this case a widely used approach consists in splitting the compositional dataset in two subsamples, one collecting the composition with structural zeros and one with all the remaining data, and assuming they have been sampled from different populations (Pawlowsky-Glahn et al., 2015). This approach suggests then to treat the two subsamples differently: in the first one only the subcompositions

of the non-null parts are retained; in the second one the whole compositions are taken in consideration.

Without loss of generality, in the remaining of the paper we will denote the two subsamples with $(\mathcal{Y}, 0)$ and $\mathcal{Z}$, respectively. The rationale behind this classical approach is that the null parts can be omitted in the subsample with structural zeros and logratio transformations can be applied to $\mathcal{Y}$. When performing dimensionality reduction with PCA, this approach will result in two independent analyses for the two subsamples. However, this may be limiting because, despite the structural zeros, the two subsamples may come from an unique population, and the distinction in two subpopulation can be misleading. From a statistical (or better geometrical) perspective, this can be represented by the situation where the two subsamples still share a common subspace that would allow to represent the whole dataset in the same geometrical space and support classification or stratification studies involving the whole population. In this paper we overcome this limitation by presenting an hypothesis testing procedure to check whether such a common subspace exists.

The paper is organized as follows. In Section 2 we revise the tools of compositional data analysis needed in our work, including the most common logratio transformations, and we thoroughly describe the classical approach for performing PCA of compositional data with and without structural zeros. In Section 3 we introduce the proposed statistical test and two techniques for approximating the null distribution of the related test statistic. The first technique is a parametric approach assuming data to be normally distributed on the simplex, whose derivation is detailed in Appendix A. The second technique is a nonparametric bootstrap method working under more general assumptions. A thorough validation of the proposed approach using simulated data is presented in Section 4 while in Section 5 we apply our approach to an experimental compositional dataset concerning respiratory microbiome measurements from two groups of patients, one undergoing antibiotic therapy and one with no treatment. Finally, our conclusions are offered in Section 6.

2. Principal component analysis with compositional data

2.1. Aitchison geometry and logratio transformations

From any vector $\mathbf{w} = (w_1, w_2, \dots, w_D) \in \mathbb{R}^D$ with positive components, we can obtain a composition by computing the *closure (to 1)* of $\mathbf{w}$, defined

as

$$\mathcal{C}(\mathbf{w}) = \mathcal{C}(w_1, w_2, \dots, w_D) = \left(\frac{w_1}{\sum_{i=1}^D w_i}, \frac{w_2}{\sum_{i=1}^D w_i}, \dots, \frac{w_D}{\sum_{i=1}^D w_i} \right) \in \mathbb{S}^D.$$

More formally, it can be showed (Pawlowsky-Glahn et al., 2015) that a D -part composition is an equivalence class in $\mathbb{R}^D$ with respect to the relationship

$$\mathbf{w} = \mathbf{u} \Leftrightarrow \mathcal{C}(\mathbf{w}) = \mathcal{C}(\mathbf{u}) \quad \mathbf{w}, \mathbf{u} \in \mathbb{R}^D.$$

Moreover, it can be proved (Aitchison, 1982) that $\mathbb{S}^D$ is an Euclidean $\mathbb{R}$ -vector space with the following operations:

- *perturbation*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \dots, x_D y_D) \quad \mathbf{x}, \mathbf{y} \in \mathbb{S}^D \quad (3)$$

- *powering*

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha) \quad \mathbf{x} \in \mathbb{S}^D, \alpha \in \mathbb{R} \quad (4)$$

- *Aitchison inner product*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \left(\ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \right) \quad \mathbf{x}, \mathbf{y} \in \mathbb{S}^D. \quad (5)$$

As mentioned in the introduction, the principle of working in coordinates suggests to perform a transformation of a compositional dataset. Here are some details on the two most commonly used. The clr transformation of the D -part composition $\mathbf{x} = (x_1, x_2, \dots, x_D)$ is defined as

$$\text{clr}(\mathbf{x}) = \left(\log \frac{x_1}{g(\mathbf{x})}, \log \frac{x_2}{g(\mathbf{x})}, \dots, \log \frac{x_D}{g(\mathbf{x})} \right),$$

where $g(\mathbf{x}) = \left(\prod_{i=1}^D x_i \right)^{1/D}$ denotes the geometric mean of $\mathbf{x}$.

The clr transformation is a function from $\mathbb{S}^D$ to $\mathbb{R}^D$, which maps an element of the simplex to a vector whose components sum to 0. Moreover, it

preserves distances and angles, meaning that the Aitchison inner product of two compositions in $\mathbb{S}^D$ is equal to the usual Euclidean inner product of the corresponding transformed vectors in $\mathbb{R}^D$ (for further details, see Pawlowsky-Glahn et al., 2015, and references within).

The ilr transformations are strictly related to the orthonormal bases of $\mathbb{S}^D$, since the selection of an orthonormal basis fully determines a specific ilr transformation (Egozcue et al., 2003). Given a Aitchison-orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$ in the simplex $\mathbb{S}^D$, that is a orthonormal basis with respect to the operations of *perturbation* and *powering* defined in Eqs. (3) and (4), the corresponding ilr transformation maps an element $\mathbf{x}$ in $\mathbb{S}^D$ to a vector in $\mathbb{R}^{D-1}$ which components are given by the coordinates of $\mathbf{x}$ with respect to that basis:

$$\text{ilr}(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, \dots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a).$$

From the definition, it clearly follows that several ilr transformations can be considered. An important one (Egozcue et al., 2003) is related to the orthonormal basis consisting of the vectors

$$\mathbf{e}_i = \mathcal{C}(\exp(\mathbf{v}_i)) \quad i = 1, 2, \dots, D-1,$$

where the vectors $\mathbf{v}_i$ are given by

$$\mathbf{v}_i = \sqrt{\frac{D-i}{D-i+1}} \left(\underbrace{0, \dots, 0}_{i-1}, 1, -\frac{1}{D-i}, \dots, -\frac{1}{D-i} \right) \quad i = 1, 2, \dots, D-1,$$

and the exponential function is component-wise computed. This basis determines to the so-called *pivot (logratio) coordinates* of the composition $\mathbf{x} = (x_1, x_2, \dots, x_D)$, denoted by

$$\tilde{\mathbf{x}} = \text{ilr}(\mathbf{x}) = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{D-1}),$$

where each component $\tilde{x}_i$ is

$$\tilde{x}_i = \sqrt{\frac{D-i}{D-i+1}} \log \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^D x_j}}, \quad i = 1, 2, \dots, D-1.$$

The inverse transformation of $\tilde{\mathbf{x}}$ to the original composition $\mathbf{x}$ is given by

$$\mathbf{x} = \text{ilr}^{-1}(\tilde{\mathbf{x}}) = \mathcal{C}(\exp(\boldsymbol{\psi})),$$

where the component of ψ are

$$\psi_j = \begin{cases} \tilde{x}_1 \sqrt{\frac{D-1}{D}} & j = 1 \\ -\sum_{i=1}^{j-1} \frac{\tilde{x}_i}{\sqrt{(D-i+1)(D-i)}} + \tilde{x}_j \sqrt{\frac{D-j}{D-j+1}} & j = 2, \dots, D-1 \\ -\sum_{i=1}^{D-1} \frac{\tilde{x}_i}{\sqrt{(D-i+1)(D-i)}} & j = D. \end{cases}$$

The relationship between the pivot (logratio) and the clr coordinates deserves to be mentioned here. It can be expressed by the following formulas:

$$\tilde{\mathbf{x}} = \text{ilr}(\mathbf{x}) = \mathbf{V}^T \text{clr}(\mathbf{x})$$

$$\mathbf{x} = \text{ilr}^{-1}(\tilde{\mathbf{x}}) = \mathcal{C}(\exp(\mathbf{V}\tilde{\mathbf{x}})),$$

being $\mathbf{V}$ a $D \times (D-1)$ matrix with entries

$$V_{ij} = \begin{cases} \frac{D-j}{\sqrt{(D-j+1)(D-j)}} & i = j \\ -1 & i > j \\ \frac{1}{\sqrt{(D-j+1)(D-j)}} & i < j \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that the following identities hold:

$$\mathbf{V}^T \mathbf{V} = \mathbf{I}_{D-1} \quad \text{and} \quad \mathbf{V} \mathbf{V}^T = \mathbf{I}_D - (1/D) \mathbf{1}_D \mathbf{1}_D^T,$$

being $\mathbf{I}_{D-1}$ the identity matrix of order D and $\mathbf{1}_D$ the column vector of $\mathbb{R}^D$ with all the components equal to one.

2.2. Aitchison's approach for PCA of compositional data

The Principal Component Analysis (PCA) is one of the first statistical methods adapted for Compositional Data Analysis. The first attempts in that direction are due to Aitchison (1983, 1984), who proposed a logratio-transformation of the data before the application of PCA, basically for two reasons. The first one is related to the marked curvature often displayed by a compositional dataset, which can not be captured by standard principal components. The second one pertains to the constant-sum constraint in

compositional data, which imposes structural restrictions on the correlation matrix of the raw proportions. Consequently, the correlations are not entirely unconstrained, and in this context, it seems of little use to insist on Euclidean orthogonality and the zero correlation of linear combinations of raw proportions (see for further details Aitchison, 1983).

Aitchison’s approach has been largely applied in many fields (Cicchella et al., 2022; Aitchison and Greenacre, 2002; Wang et al., 2015). Effectively, it has revealed itself more adequate in capturing the non-linear curved patterns often displayed by compositional datasets than classical PCA performed on the original data (see for example Aitchison, 1983; Aitchison and Greenacre, 2002; Filzmoser et al., 2009, 2018). In this paper, following the approach suggested by Filzmoser et al. (2009), we apply Aitchison’s approach by considering the ilr transformation, since it allows to obtain non-collinear data, and thus full-rank covariance matrices of the data.

A relevant drawback of such a procedure is that it cannot be applied whenever there are data with value zero, because the argument of the logarithm function must be strictly positive. As mentioned in the previous section, this issue may be overcome by replacing the zeros by a (small) positive value and then applying the logratio-transformation to the modified dataset. Nevertheless, in the presence of many zero parts, and especially in the case of structural zeros, the replacement cannot be considered a reasonable choice nor a good practice.

2.3. PCA of compositional data with structural zeros

This section addresses the issue about how performing a PCA in presence of structural zeros in the data. First we need to introduce the setting that we will use throughout the paper.

In a compositional dataset as described in Eq. (2), we assume that n_y observations (i.e. compositions) have Q parts (with $Q < D$) that are structural zeros. Without loss of generality we can assume

$$\begin{aligned} \mathcal{X} &= \begin{pmatrix} \mathcal{Y} & \mathbf{0}_{n_y, Q} \\ \mathcal{Z} & \end{pmatrix} = \begin{pmatrix} (\mathbf{y}_1, \dots, \mathbf{y}_{n_y})' & \mathbf{0}_{n_y, Q} \\ (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{n_z})' & \end{pmatrix}, \text{ where:} \\ \mathbf{y}_i &= (y_{i1}, \dots, y_{i(D-Q)}) \in \mathbb{S}^{(D-Q)}, \quad \sum_{j=1}^{D-Q} y_{ij} = 1 \quad i = 1, 2, \dots, n_y \\ \mathbf{z}_i &= (z_{i1}, \dots, z_{iD}) \in \mathbb{S}^D, \quad \sum_{j=1}^D z_{ij} = 1 \quad i = 1, 2, \dots, n_z. \end{aligned} \quad (6)$$

Moreover, in this setting, $n_y + n_z = n$ and $\mathbf{0}_{n_y, Q}$ is a matrix of size $n_y \times Q$ whose elements are all equal to zero.

In the presence of structural zeros, logratio transformations cannot be applied to the data, and thus Aitchinson’s approach based on performing a PCA on the logratio-transformed data cannot be used. The current available methods for facing this issue can be grouped in three approaches.

The first approach consists in the replacement of the zeros by an arbitrary small value. This procedure is very easy and effective, but, as argued in Greenacre (2018, p. 57-58), regardless of the zero-replacement strategy employed, the introduction of new values modifies the row totals of the dataset, thereby violating the constant-sum constraint. Then, it becomes necessary to apply the closure to the rows. These modifications in the data, however, may significantly influence the results of the analyses.

A second approach suggests dropping out the zeros and performing the analyses with the remaining subcompositions with non-zero parts. This approach is also very easy to implement, but it has a relevant drawback: it is evident that the removal of all the null parts affects the amount of information conveyed by the data. In some cases, this loss of information can be significant, making the results of the analyses unreliable.

Finally, the third approach consists in applying a different transformation, avoiding the logratios (see, for example, Scealy et al. (2015), Lu et al. (2024) and reference therein). Among the transformations described in the literature, introduced for example in Tsagris et al. (2011) or in Greenacre (2024), perhaps the most common one is the square root. The calculation of the square root of each part can be easily executed also in case of zero values, and it has the appealing peculiarity to modify the dataset into a directional dataset, which can be handled by appropriate tools belonging to the field of directional statistics (Fisher, 1993; Cuesta-Albertos et al., 2009; Lee, 2010; Pewsey et al., 2013; Pewsey and García-Portugués, 2021; Porro et al., 2024). Although this method can provide interesting and useful results, it does not adhere to Aitchison’s original spirit (Alenazi, 2021).

3. Proposed test on the common subspace

Following the spirit of Aitchison’s approach in presence of structural zeros, the procedure to reduce the dimension of the dataset by performing a PCA should consist in splitting the two sub-datasets $\mathcal{Y}$ and $\mathcal{Z}$ and executing

two PCAs with $D - Q + 1$ and D principal components, respectively. The rationale of this approach is that in the Aitchison framework, the two sub-datasets are the representation of two different subpopulations: one related to the observations with structural zeros and one related to the remaining observations (without structural zeros). We follow this approach, and we try to overcome a relevant constraint: the observations can come from two subpopulations (identifiable from the presence or the absence of the structural zeros), but in many real cases, the investigation on what they share can be relevant. It means to understand whether the first (and more important) principal components of the two datasets span the same space. If this is true, it can be interpreted as evidence that both the subpopulations actually share common characteristics; therefore, they can be considered as parts of a whole and unique population.

Inspired by Aitchison's approach, we apply a PCA to the logratio-transformed data including in the sub-datasets $\mathcal{Y}$ and $\mathcal{Z}$. First we compute the transformed dataset

$$\tilde{\mathcal{X}} = \begin{pmatrix} \tilde{\mathcal{Y}} & \mathbf{0}_{n_y, Q} \\ \tilde{\mathcal{Z}} & \end{pmatrix} = \begin{pmatrix} (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_{n_y})' & \mathbf{0}_{n_y, Q} \\ (\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_{n_z})' & \end{pmatrix} \quad (7)$$

where $\tilde{\mathbf{y}}_i = \text{ilr}(\mathbf{y}_i)$ for all $i = 1, \dots, n_y$ and $\tilde{\mathbf{z}}_i = \text{ilr}(\mathbf{z}_i)$ for all $i = 1, \dots, n_z$. Then, we consider the sample covariance matrices $\hat{\boldsymbol{\Omega}}_{\mathbf{Y}}$ and $\hat{\boldsymbol{\Omega}}_{\mathbf{Z}}$ of $\tilde{\mathcal{Y}}$ and $\tilde{\mathcal{Z}}$, respectively, and we compute the eigenvalue decompositions

$$\hat{\boldsymbol{\Omega}}_{\mathbf{Y}} = \hat{\mathbf{U}} \text{diag}(\hat{\alpha}_1, \dots, \hat{\alpha}_{D-Q-1}) \hat{\mathbf{U}}' \quad \hat{\alpha}_1 > \dots > \hat{\alpha}_{D-Q-1} \quad (8)$$

and

$$\hat{\boldsymbol{\Omega}}_{\mathbf{Z}} = \hat{\mathbf{V}} \text{diag}(\hat{\beta}_1, \dots, \hat{\beta}_{D-1}) \hat{\mathbf{V}}' \quad \hat{\beta}_1 > \dots > \hat{\beta}_{D-1}. \quad (9)$$

We observe that if the first K principal components in Eqs. (8) and (9) span the same subspace then this approach can be easily used for dimensionality reduction of the whole dataset $\tilde{\mathcal{X}}$. The main difficulty in comparing these components is that they belong to different vector spaces, because the columns of $\hat{\mathbf{V}}$ and $\hat{\mathbf{U}}$ are vectors of $\mathbb{R}^{D-1}$ and $\mathbb{R}^{D-Q-1}$, respectively. This issue may be overcome by applying the canonical inclusion that transforms $\hat{\mathbf{u}}_i \in \mathbb{R}^{D-Q-1}$ in $\begin{pmatrix} \hat{\mathbf{u}}_i \\ \mathbf{0}_{Q,1} \end{pmatrix} \in \mathbb{R}^{D-1}$. Motivated by this consideration, we introduce the following definition, which extends the concept of common principal component subspace proposed by Schott (1988) to settings where a group of variables have structural zeros.

Definition 1. For all $j = 1, 2$, let $\mathcal{H}^{(j)} \in \mathbb{R}^{n^{(j)} \times D^{(j)}}$ be a dataset collecting $n^{(j)}$ observations of a $D^{(j)}$ -dimensional random vector with covariance matrix $\mathbf{\Omega}^{(j)}$, being $D^{(1)} < D^{(2)}$. Fixed $K \in \{1, \dots, D^{(1)}\}$, we say that $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ share a common principal component subspace of size K if

$$\text{span} \left(\begin{pmatrix} \mathbf{w}_1^{(1)} \\ \mathbf{0}_{Q,1} \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{w}_K^{(1)} \\ \mathbf{0}_{Q,1} \end{pmatrix} \right) = \text{span}(\mathbf{w}_1^{(2)}, \dots, \mathbf{w}_K^{(2)}) \quad (10)$$

where $Q = D^{(2)} - D^{(1)}$ while $\mathbf{w}_1^{(1)}, \dots, \mathbf{w}_K^{(1)}$ and $\mathbf{w}_1^{(2)}, \dots, \mathbf{w}_K^{(2)}$ are the eigenvectors associated to the K highest eigenvalues of $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$, respectively.

Definition 2. Let $\mathcal{X} = \begin{pmatrix} \mathcal{Y} & \mathbf{0}_{n_y, Q} \\ \mathcal{Z} & \end{pmatrix}$ be a compositional dataset with structural zeros as in Eq. (6). Fixed $K \in \{1, \dots, D - Q - 1\}$, we say that $\mathcal{Y}$ and $\mathcal{Z}$ share a common principal component subspace of size K if the corresponding dataset in pivot logratio coordinates $\tilde{\mathcal{Y}}$ and $\tilde{\mathcal{Z}}$ share a common principal component subspace of size K .

Following Definitions 1 and 2, the main objective of this paper is to develop a statistical procedure for testing the null hypothesis

$$\mathbf{H}_0 : \text{span} \left(\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{0}_{Q,1} \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{u}_K \\ \mathbf{0}_{Q,1} \end{pmatrix} \right) = \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_K) \quad (11)$$

against the alternative hypothesis

$$\mathbf{H}_1 : \text{span} \left(\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{0}_{Q,1} \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{u}_K \\ \mathbf{0}_{Q,1} \end{pmatrix} \right) \neq \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_K) \quad (12)$$

where $\mathbf{u}_j \in \mathbb{R}^{D-Q-1}$ and $\mathbf{v}_j \in \mathbb{R}^{D-1}$ denote the j -th principal component (PC) of $\tilde{\mathcal{Y}}$ and $\tilde{\mathcal{Z}}$ respectively.

Towards this end, inspired by Schott (1988), we define the following test statistic, whose computation is summarized in Algorithm 1.

Definition 3. Let $\mathcal{X}$ a compositional dataset with structural zeros as in Eq. (6). Assume that $\mathcal{Y}$ and $\mathcal{Z}$ are realization of random samples $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_y} \sim \mathbf{Y}$ and $\mathbf{Z}_1, \dots, \mathbf{Z}_{n_z} \sim \mathbf{Z}$, where $\mathbf{Y}$ is a $(D - Q)$ -part random composition and $\mathbf{Z}$ is a D -part random composition. Fixed $K \in \{1, \dots, D - Q\}$ and denoted with $\hat{\mathbf{\Omega}}_{\mathbf{Y}}$ and $\hat{\mathbf{\Omega}}_{\mathbf{Z}}$ the sample covariance matrix of the data in pivot logratio coordinates, we define the test statistic

$$T := \sum_{i=1}^K \left((n_y - 1) \hat{\alpha}_i + (n_z - 1) \hat{\beta}_i - \hat{\gamma}_i \right) \quad (13)$$

where $\hat{\alpha}_i$, $\hat{\beta}_i$, and $\hat{\gamma}_i$ denote the i -th largest eigenvalue of $\hat{\Omega}_{\mathbf{Y}}$, $\hat{\Omega}_{\mathbf{Z}}$, and

$$(n_y - 1) \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-1)} & \end{bmatrix} + (n_z - 1) \hat{\Omega}_{\mathbf{Z}}, \quad (14)$$

respectively.

Algorithm 1 Computation of the test statistic

Input: Compositional dataset $\mathcal{X} = \begin{pmatrix} \mathcal{Y} & \mathbf{0}_{n_y,Q} \\ \mathcal{Z} & \end{pmatrix}$ with structural zeros as

in Eq. (6); $K \in \{1, \dots, D - Q\}$;

1: **for** $i = 1, \dots, n_y$ **do**

2: $\tilde{\mathbf{y}}_i \leftarrow \text{ilr}(\mathbf{y}_i)$

3: **end for**

4: **for** $i = 1, \dots, n_z$ **do**

5: $\tilde{\mathbf{z}}_i \leftarrow \text{ilr}(\mathbf{z}_i)$

6: **end for**

7: Assemble $\tilde{\mathcal{Y}} = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_{n_y})'$ and $\tilde{\mathcal{Z}} = (\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_{n_z})'$

8: Compute sample covariance matrices $\hat{\Omega}_{\mathbf{Y}}$ and $\hat{\Omega}_{\mathbf{Z}}$

9: Compute K highest eigenvalues of $\hat{\Omega}_{\mathbf{Y}}$: $\hat{\alpha}_1 \geq \dots \geq \hat{\alpha}_K$

10: Compute K highest eigenvalues of $\hat{\Omega}_{\mathbf{Z}}$: $\hat{\beta}_1 \geq \dots \geq \hat{\beta}_K$

11: Compute K highest eigenvalues of $(n_y - 1) \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{D-Q,Q} \\ \mathbf{0}_{Q,D} & \end{bmatrix} + (n_z - 1) \hat{\Omega}_{\mathbf{Z}}$:

$\hat{\gamma}_1 \geq \dots \geq \hat{\gamma}_K$

12: Compute $t = \sum_{j=1}^K \left((n_y - 1) \hat{\alpha}_j + (n_z - 1) \hat{\beta}_j - \hat{\gamma}_j \right)$

Output: Test statistic t

3.1. Approximate test for normally distributed data

Suppose that $\mathbf{y}_1, \dots, \mathbf{y}_{n_y}$ are realizations of a $(D - Q)$ -part random composition $\mathbf{Y}$ having a normal distribution on the simplex $\mathbb{S}^{D-Q}$ with ilr-mean $\boldsymbol{\mu}_{\mathbf{Y}}$ and ilr-covariance matrix $\boldsymbol{\Omega}_{\mathbf{Y}}$, namely $\mathbf{Y} \sim \mathcal{N}_{\mathbb{S}^{D-Q}}(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Omega}_{\mathbf{Y}})$. It follows by definition (Egozcue and Pawlowsky-Glahn, 2019) that the corresponding pivot logratio coordinates are realization of a $(D - Q - 1)$ -dimensional random vector $\tilde{\mathbf{Y}} := \text{ilr}(\mathbf{Y})$ with multivariate normal distribution $\mathcal{N}_{\mathbb{R}^{D-Q-1}}(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Omega}_{\mathbf{Y}})$. Analogously, suppose that $\mathbf{z}_1, \dots, \mathbf{z}_{n_z}$ are realizations of a D -part random composition $\mathbf{Z} \sim \mathcal{N}_{\mathbb{S}^D}(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Omega}_{\mathbf{Z}})$, and thus $\tilde{\mathbf{Z}} := \text{ilr}(\mathbf{Z}) \sim \mathcal{N}_{\mathbb{R}^{D-1}}(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Omega}_{\mathbf{Z}})$.

In this scenario, by exploiting that both $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Z}}$ have a multivariate normal distribution, we are able to derive a analytical approximation for the distribution of the test statistic T introduced in Definition 3 under the null hypothesis that $\mathcal{Y}$ and $\mathcal{Z}$ share a common principal subspace, see Eq. (11). This result is summarized in the next theorem and is a generalization of the procedure proposed by Schott (1988) to the case in which the PCs of the two datasets belong to vector spaces with different dimensions.

Theorem 1. *Let us assume that the hypothesis of Definition 3 hold and that $\mathbf{Y} \sim \mathcal{N}_{\mathbb{S}^{D-Q}}(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Omega}_{\mathbf{Y}})$ and $\mathbf{Z} \sim \mathcal{N}_{\mathbb{S}^D}(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Omega}_{\mathbf{Z}})$. Then, the distribution of T under the null hypothesis $\mathbf{H}_0$ in Eq. (11) can be approximated as*

$$T \sim \frac{\sigma_T^2}{2\mu_T} \chi^2 \left(\left\lceil \frac{2\mu_T^2}{\sigma_T^2} \right\rceil \right), \quad (15)$$

where

$$\begin{aligned} \mu_T = & \sum_{i=1}^K \sum_{j=K+1}^{D-1} \left\{ \frac{\alpha_i \alpha_j}{(\alpha_i - \alpha_j)} + \frac{\beta_i \beta_j}{(\beta_i - \beta_j)} + \right. \\ & \left. - \frac{\sum_{h=1}^K \sum_{l=K+1}^{D-1} [(n_y - 1)(u_{ih}^*)^2 (u_{jl}^*)^2 \alpha_h \alpha_l + (n_z - 1)(v_{ih}^*)^2 (v_{jl}^*)^2 \beta_h \beta_l]}{(n_y + n_z - 2)(\psi_i - \psi_j)} \right\} \end{aligned} \quad (16)$$

and

$$\begin{aligned} \sigma_T^2 = & 2 \sum_{i=1}^K \sum_{j=K+1}^{D-1} \left\{ \frac{\alpha_i^2 \alpha_j^2}{(\alpha_i - \alpha_j)^2} + \frac{\beta_i^2 \beta_j^2}{(\beta_i - \beta_j)^2} - \frac{2}{(n_y + n_z - 2)(\psi_i - \psi_j)} \times \right. \\ & \times \sum_{h=1}^K \sum_{l=K+1}^{D-1} \left[\frac{(n_y - 1)(\alpha_h \alpha_l u_{ih}^* u_{jl}^*)^2}{(\alpha_h - \alpha_l)} + \frac{(n_z - 1)(\beta_h \beta_l v_{ih}^* v_{jl}^*)^2}{(\beta_h - \beta_l)} \right] + \\ & \left. + \sum_{h=1}^K \sum_{l=K+1}^{D-1} \frac{W_{hl}^2}{(n_y + n_z - 2)^2 (\psi_i - \psi_j)(\psi_h - \psi_l)} \right\}. \end{aligned} \quad (17)$$

with

$$\begin{aligned} W_{hl} = & (n_y - 1) \left(\sum_{s=1}^K \alpha_s u_{is}^* u_{hs}^* \right) \left(\sum_{t=K+1}^{D-1} \alpha_t u_{jt}^* u_{lt}^* \right) \\ & + (n_z - 1) \left(\sum_{s=1}^K \beta_s v_{is}^* v_{hs}^* \right) \left(\sum_{t=K+1}^{D-1} \beta_t v_{jt}^* v_{lt}^* \right). \end{aligned}$$

In Eqs. (16) and (17), $\alpha_1, \dots, \alpha_{D-Q-1}$, $\beta_1, \dots, \beta_{D-1}$, and $\psi_1, \dots, \psi_{D-1}$ denote the eigenvalues of $\mathbf{\Omega}_{\mathbf{Y}}$, $\mathbf{\Omega}_{\mathbf{Z}}$, and of the pooled covariance matrix

$$\mathbf{\Omega}_{pool} = (n_y + n_z - 2)^{-1} \left[(n_y - 1) \begin{bmatrix} \mathbf{\Omega}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-1)} & \end{bmatrix} + (n_z - 1) \mathbf{\Omega}_{\mathbf{Z}} \right], \quad (18)$$

respectively, while $\alpha_{D-Q} = \dots = \alpha_{D-1} = 0$.

Furthermore, denoted with $\mathbf{U} \in \mathbb{R}^{(D-Q-1) \times (D-Q-1)}$, $\mathbf{V} \in \mathbb{R}^{(D-1) \times (D-1)}$, and $\mathbf{K} \in \mathbb{R}^{(D-1) \times (D-1)}$ the matrix of the normalized eigenvectors of $\mathbf{\Omega}_{\mathbf{Y}}$, $\mathbf{\Omega}_{\mathbf{Z}}$, and $\mathbf{\Omega}_{pool}$, respectively, we defined $\mathbf{U}^* = \mathbf{K}' \begin{bmatrix} \mathbf{U} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-Q-1)} & \mathbf{I}_Q \end{bmatrix}$ and $\mathbf{V}^* = \mathbf{K}' \mathbf{V}$.

The detailed proof of Theorem 1 can be found in Appendix A.2.

Eq. (16) and (17) describe the parameters μ_T and σ_T^2 as functions of the eigenvectors and eigenvalues of the covariance matrices $\mathbf{\Omega}_{\mathbf{Y}}$ and $\mathbf{\Omega}_{\mathbf{Z}}$ that are usually unknowns. To make Theorem 1 applicable, we need, therefore, an estimate of such quantities. The unknown values can be estimated using the sample covariance matrices in pivot logratio coordinates as shown in the Algorithm 2.

Algorithm 2 Approximation of the parameters of the null distribution for normally distributed data.

Input: ilr-transformed dataset $\tilde{\mathcal{X}} = \begin{pmatrix} \tilde{\mathcal{Y}} & \mathbf{0}_{n_y, Q} \\ \tilde{\mathcal{Z}} \end{pmatrix}$ as in Eq. (7)

- 1: Compute $\hat{\Omega}_{\mathbf{Y}}$, eigenvectors $\hat{\mathbf{U}}$ and eigenvalues $\hat{\alpha}_1 \geq \dots \geq \hat{\alpha}_{D-Q-1}$
- 2: Compute $\hat{\Omega}_{\mathbf{Z}}$, eigenvectors $\hat{\mathbf{V}}$ and eigenvalues $\hat{\beta}_1 \geq \dots \geq \hat{\beta}_{D-1}$
- 3: Compute matrix in Eq. (14), eigenvectors $\hat{\mathbf{K}}$ and eigenvalues $\hat{\gamma}_1 \geq \dots \geq \hat{\gamma}_{D-1}$
- 4: Decompose $\hat{\mathbf{K}} = [\hat{\mathbf{K}}_1, \hat{\mathbf{K}}_2]$, $\hat{\mathbf{K}}_1 \in \mathbb{R}^{(D-1) \times K}$, $\hat{\mathbf{K}}_2 \in \mathbb{R}^{(D-1) \times (D-K-1)}$
- 5: Compute $\hat{\mathbf{U}}_1^* = (\hat{u}_{ij}^*)_{i,j=1,\dots,K}$ and $\hat{\mathbf{U}}_2^* = (\hat{u}_{ij}^*)_{i,j=K+1,\dots,D-1}$ as the eigenvectors of $(n_y - 1)\hat{\mathbf{K}}_1' \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{D-Q, Q} \\ \mathbf{0}_{Q, D} \end{bmatrix} \hat{\mathbf{K}}_1$ and $(n_y - 1)\hat{\mathbf{K}}_2' \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{D-Q, Q} \\ \mathbf{0}_{Q, D} \end{bmatrix} \hat{\mathbf{K}}_2$
- 6: Compute $\hat{\mathbf{V}}_1^* = (\hat{v}_{ij}^*)_{i,j=1,\dots,K}$ and $\hat{\mathbf{V}}_2^* = (\hat{v}_{ij}^*)_{i,j=K+1,\dots,D-1}$ as the eigenvectors of $(n_z - 1)\hat{\mathbf{K}}_1' \hat{\Omega}_{\mathbf{Z}} \hat{\mathbf{K}}_1$ and $(n_z - 1)\hat{\mathbf{K}}_2' \hat{\Omega}_{\mathbf{Z}} \hat{\mathbf{K}}_2$
- 7: Approximated $\hat{\mu}_T$ and $\hat{\sigma}_T^2$ through Eq. (16) and (17) by replacing:
 $\alpha_i \rightarrow \hat{\alpha}_i, i = 1, \dots, D - Q - 1$; $\alpha_i \rightarrow 0, i = D - Q, \dots, D - 1$
 $\beta_i \rightarrow \hat{\beta}_i$; $\psi_i \rightarrow \hat{\gamma}_i / (n_y + n_z - 2)$; $u_{ij}^* \rightarrow \hat{u}_{ij}^*$; $v_{ij}^* \rightarrow \hat{v}_{ij}^*$

Output: $\hat{\mu}_T, \hat{\sigma}_T^2$

3.2. Nonparametric bootstrap test

The adapted Schott's procedure described in the previous section allows an analytic approximation of the null distribution of the test statistic T , provided that data are normally distributed on the simplex. Here we discuss a nonparametric approach to be applied when such an approximation does not hold. Since the presence of structural zeros prevents the permutation of observations between the two datasets $\mathcal{Y}$ and $\mathcal{Z}$, we rely our approach on the bootstrap method described in Algorithm 3.

The Monte Carlo approximation of the distribution of the test statistic under the null hypothesis, and thus the approximation of the p-value, is performed as follows. We generate pairs of datasets $(\mathcal{Y}^b, \mathcal{Z}^b)$ that satisfy the null hypothesis in Eq. (11), i.e. share a common subspace of dimension K , even though the corresponding principal components are different. To this end, we start from the sample covariance matrix $\hat{\Omega}_{\mathbf{Y}}$ of the ilr-transformed

sub-dataset with structural zeros and we randomly rotate the subspace in $\mathbb{R}^{D-1}$ generated by its first K principal components and its orthogonal complement, that is we define

$$\hat{\mathbf{U}}^{boot} = \begin{bmatrix} \hat{\mathbf{U}} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-Q)-1} & \mathbf{I}_Q \end{bmatrix} \begin{bmatrix} \mathbf{R}_1^{boot} & \mathbf{0}_{K,(D-K-1)} \\ \mathbf{0}_{(D-K-1),K} & \mathbf{R}_2^{boot} \end{bmatrix} \quad (19)$$

where $\mathbf{R}_1^{boot}$ and $\mathbf{R}_2^{boot}$ are random rotation matrices of size $K \times K$ and $(D-K-1) \times (D-K-1)$, respectively. We then apply to the dataset $\tilde{\mathbf{Z}}$ the rotation $\mathbf{R}_3^{boot} = \hat{\mathbf{U}}^{boot} \hat{\mathbf{V}}'$, that aligns its principal components to $\hat{\mathbf{U}}^{boot}$. Denoted with $\tilde{\mathbf{Z}}^{boot}$ this rotated sub-dataset, the bootstrap datasets $\mathcal{Y}^b$ and $\mathcal{Z}^b$ are formed by drawing with replacement n_y samples from $\tilde{\mathbf{Z}}$ and n_z samples from $\tilde{\mathbf{Z}}^{boot}$, respectively.

Algorithm 3 Nonparametric bootstrap test

Input: ilr-transformed dataset $\tilde{\mathcal{X}} = \begin{pmatrix} \tilde{\mathcal{Y}} & \mathbf{0}_{n_y,Q} \\ \tilde{\mathcal{Z}} & \end{pmatrix}$ as in Eq. (7)

$\hat{\mathbf{\Omega}}_{\mathbf{Y}}$ and $\hat{\mathbf{\Omega}}_{\mathbf{Z}}$, and corresponding eigenvectors $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$

Sample value of the test statistic t

$n_{boot}; K$

- 1: Randomly generate orthonormal matrices $\mathbf{R}_1^{boot} \in \mathbb{R}^{K \times K}$ and $\mathbf{R}_2^{boot} \in \mathbb{R}^{(D-K-1) \times (D-K-1)}$
- 2: Define $\hat{\mathbf{U}}^{boot} = \begin{bmatrix} \hat{\mathbf{U}} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-Q)-1} & \mathbf{I}_Q \end{bmatrix} \begin{bmatrix} \mathbf{R}_1^{boot} & \mathbf{0}_{K,(D-K-1)} \\ \mathbf{0}_{(D-K-1),K} & \mathbf{R}_2^{boot} \end{bmatrix}$
- 3: Define $\mathbf{R}_3^{boot} = \hat{\mathbf{U}}^{boot} \hat{\mathbf{V}}'$
- 4: Define $\tilde{\mathbf{Z}}^{boot} = \tilde{\mathbf{Z}} (\mathbf{R}_3^{boot})'$
- 5: **for** $b = 1, \dots, n_{boot}$ **do**
- 6: Define $\tilde{\mathcal{Y}}^b$ by drawing with repetition n_y samples from $\tilde{\mathcal{Y}}$
- 7: Define $\tilde{\mathcal{Z}}^b$ by drawing with repetition n_z samples from $\tilde{\mathcal{Z}}^{boot}$
- 8: Compute the sample test statistic t^b from $\tilde{\mathcal{X}}^b = \begin{pmatrix} \tilde{\mathcal{Y}}^b & \mathbf{0}_{n_y,Q} \\ \tilde{\mathcal{Z}}^b & \end{pmatrix}$ as in

Algorithm 1

9: **end for**

10: Approximate the p-value $p^{boot} = \frac{\#\{t^b \geq t\}}{n_{boot}}$

Output: p^{boot}

4. Simulation study

In this section we use simulated data to demonstrate the validity of the technique introduced in the previous sections. Towards this end, we considered datasets formed by compositions of $D = 8$ parts, where the first n_y observations have $Q = 2$ parts that are structural zeros. We then set $K = 2$ and we tested the null hypothesis in Eq. (11) in three different scenarios.

- S1.* We simulated data under the null hypothesis by assuming that the first two PCs for the ilr-transformed datasets $\tilde{\mathcal{Y}}$ and $\tilde{\mathcal{Z}}$ span the same subspace. To this end we define the covariance matrices

$$\mathbf{\Omega}_{\mathbf{Y}} = \mathbf{U} \text{diag}(\alpha_1, \dots, \alpha_5) \mathbf{U}' \quad (20)$$

$$\mathbf{\Omega}_{\mathbf{Z}} = \mathbf{V} \text{diag}(\beta_1, \dots, \beta_7) \mathbf{V}' \quad (21)$$

where $\mathbf{U}$ is a randomly generated orthonormal matrix (Stewart, 1980; Mezzadri, 2007), and

$$\mathbf{V} = \begin{bmatrix} \mathbf{U} & \mathbf{0}_{5,2} \\ \mathbf{0}_{2,5} & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 & \mathbf{0}_{2,5} \\ \mathbf{0}_{5,2} & \mathbf{R}_2 \end{bmatrix} \quad (22)$$

being $\mathbf{R}_1 \in \mathbb{R}^{2 \times 2}$ and $\mathbf{R}_2 \in \mathbb{R}^{5 \times 5}$ random rotation matrices.

- S2.* Only the first PC is common for the ilr-transformed datasets $\tilde{\mathcal{Y}}$ and $\tilde{\mathcal{Z}}$. As in the previous scenario, we randomly sampled $\mathbf{U}$, while we defined

$$\mathbf{V} = \begin{bmatrix} \mathbf{U} & \mathbf{0}_{5,2} \\ \mathbf{0}_{2,5} & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}_{1,6} \\ \mathbf{0}_{6,1} & \mathbf{R} \end{bmatrix} \quad (23)$$

being $\mathbf{R} \in \mathbb{R}^{6 \times 6}$ a random rotation matrix.

- S3.* The ilr-transformed datasets $\tilde{\mathcal{Y}}$ and $\tilde{\mathcal{Z}}$ do not share any common structure. In this case we defined $\mathbf{U}$ and $\mathbf{V}$ as independent randomly generated orthonormal matrices.

In each scenario, we sample the ilr-transformed data from three different families of probability distributions:

1. Zero-mean multivariate normal distributions with covariance matrices $\mathbf{\Omega}_{\mathbf{Y}}$ and $\mathbf{\Omega}_{\mathbf{Z}}$;

2. Zero-mean multivariate Student's t -distribution with scale matrices $\frac{\nu-2}{\nu}\mathbf{\Omega}_Y$ and $\frac{\nu-2}{\nu}\mathbf{\Omega}_Z$, where ν is the number of degrees of freedom. In the simulation below we set $\nu \in \{4, 8, 40\}$;
3. Uniform distributions on the hypercubes $[-\sqrt{3}, \sqrt{3}]^5$ and $[-\sqrt{3}, \sqrt{3}]^7$ rotated through the Cholesky decomposition of $\mathbf{\Omega}_Y$ and $\mathbf{\Omega}_Z$.

With the choices above we obtain perfectly comparable results, since all the initial marginal distributions are standardized.

4.1. Approximation of the null distribution

Our first experiment aimed at assessing our generalization of Schott's approximation of the null distribution of the test statistic under different distributions for the input data. To this end, we set $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (10, 9, 1, 1, 0.5)$ and $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7) = (6, 5, 1, 0.9, 0.3, 0.1, 0.02)$ and we generated $\mathbf{\Omega}_Y$ and $\mathbf{\Omega}_Z$ according to scenario *S1*. Then for each probability distribution described in the previous section we sampled 1000 balanced datasets so that $n_y = n_z = 100$.

Figure 1 shows the empirical cumulative distribution function (cdf) of the test statistic T computed from the simulated data and the cdf associated to our adapted Schott's approximation of the null distribution computed by exploiting knowledge of the true covariance matrices as described in Section 3.1, Eqs. (16) and (17). When the simulated data are normally distributed the empirical cdf of the test statistic resembles the one predicted using our adapted Schott's approximation, hence supporting its reliability. When a uniform distribution is used for generating the data, the empirical cdf lies consistently to the left of the one predicted through Schott's approximation, suggesting that a statistical test based on such an approximation may turn out to be too conservative. Conversely, when a multivariate t distribution underlines the simulated data the statistical test based on Schott's approximation may results too liberal as the empirical cdf tends to be on the right of that predicted using Schott's approximation. However, as expected, the distance between empirical and predicted cdf decreases for increasing value of the degree of freedom.

4.2. Power analysis

For each scenario and each distribution we carried out 1000 simulations testing different sample sizes, namely $n_y \in \{20, 60, 100\}$ and $n_z \in \{20, 60, 100\}$.

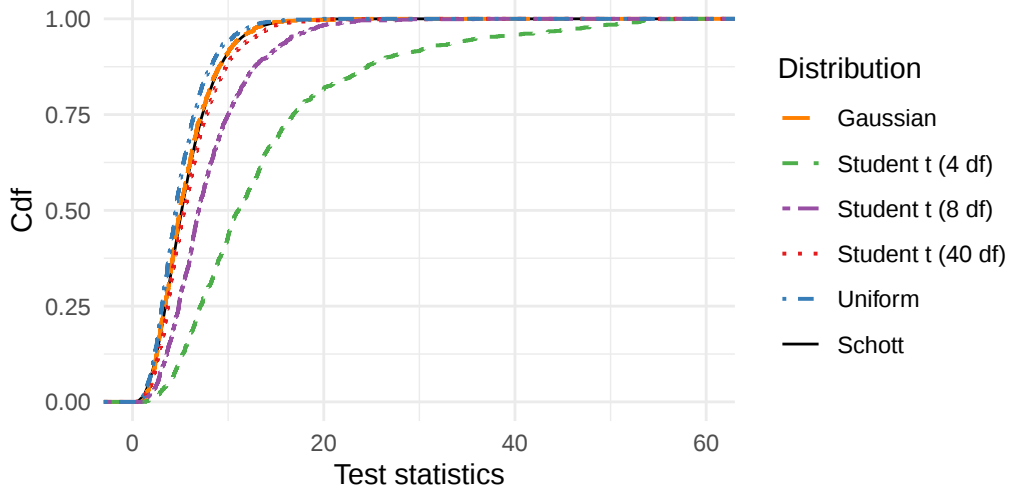


Figure 1: Cumulative distribution function (cdf) associated to the empirical null distribution of the test statistic T computed by simulating datasets with different multivariate distributions (colored lines). As a touchstone, we plotted the cdf of the approximation of the null distribution obtained by generalizing Schott’s formula (black solid line)

The eigenvalues of the covariance matrices were set as described in the previous section. Fixed $K = 2$, we tested the null hypothesis in Eq. (11) against the alternative hypothesis in Eq. (12) through the three procedures described in the previous section, i.e. by approximating the null distribution of the test statistic using Schott’s formula with the true covariance matrices (*Schott theo.*) and with the sample ones (*Schott est.*) and using the bootstrap procedure described in Algorithm 3 (*Bootstrap*). In the *Bootstrap* test, $n_{boot} = 1000$ bootstrap samples were drawn. Of course, the *Schott theo.* results can only be obtained in a simulation framework, since the true covariance matrix is not known in real data analyses. However, in this section we present also the *Schott theo.* results, as some comparisons between *Schott theo.* and *Schott est.* are worth discussing.

To compare the three tests, we set their nominal level equal to 5% and we estimated the probability of rejected $\mathbf{H}_0$ that corresponds to the probability of type I error in Scenario *S1* and with the power of the test in scenarios *S2* and *S3*.

Under Gaussian assumption for the simulated data, *Schott theo.* exhibits in most cases the type I error rate closest to the nominal value (Table 1)

		(n_y, n_z)								
		(20,20)	(20,60)	(20,100)	(60,20)	(60,60)	(60,100)	(100,20)	(100,60)	(100,100)
Gaussian	Schott theo.	0.068	0.043	0.057	0.064	0.053	0.047	0.054	0.054	0.054
	Schott est.	0.095	0.063	0.081	0.102	0.074	0.048	0.095	0.066	0.063
	Bootstrap	0.079	0.060	0.089	0.101	0.073	0.051	0.095	0.069	0.069
Student's t (40 df)	Schott theo.	0.060	0.080	0.074	0.074	0.069	0.070	0.066	0.079	0.069
	Schott est.	0.081	0.106	0.097	0.103	0.077	0.076	0.101	0.084	0.078
	Bootstrap	0.062	0.093	0.088	0.088	0.059	0.064	0.094	0.071	0.071
Student's t (8 df)	Schott theo.	0.154	0.151	0.166	0.145	0.182	0.172	0.150	0.177	0.187
	Schott est.	0.188	0.177	0.201	0.198	0.190	0.181	0.195	0.189	0.201
	Bootstrap	0.108	0.122	0.112	0.130	0.095	0.081	0.137	0.084	0.077
Student's t (4 df)	Schott theo.	0.248	0.352	0.331	0.273	0.424	0.512	0.309	0.451	0.539
	Schott est.	0.351	0.436	0.446	0.373	0.471	0.550	0.411	0.485	0.573
	Bootstrap	0.176	0.189	0.174	0.193	0.144	0.138	0.237	0.128	0.150
Uniform	Schott theo.	0.040	0.022	0.031	0.031	0.034	0.026	0.030	0.039	0.032
	Schott est.	0.068	0.038	0.060	0.057	0.045	0.029	0.058	0.039	0.033
	Bootstrap	0.069	0.056	0.079	0.078	0.069	0.054	0.071	0.054	0.062

Table 1: Estimated probability of type I error for varying sample sizes and sampling probability. Data were simulated according to the null hypothesis (Scenario $S1$). The nominal value of tests is 5%.

and the highest power (Tables 2 and 3). Instead, *Schott est.* and *Bootstrap* provide comparable results, although *Bootstrap* outperforms *Schott est.* for small sample sizes.

In accordance with the analysis of the previous section, when data are sampled from multivariate Student's t -distributions all tests tends to be too liberal. However, in this scenario the *Bootstrap* test shows in general a lower probability of type I error, see Table 1. To highlight the most extreme example, when the number of degree of freedom is 4, the estimated probability of type I error ranges from 0.128 to 0.237 for the *Bootstrap* test, from 0.248 to 0.539 for *Schott theo.*, and from 0.573 to 0.351 for *Schott est.*. Moreover, the power of the *Bootstrap* test is usually higher than the power of *Schott est.*, see Tables 2 and 3.

When a multivariate Uniform distribution is used to simulate data, the tests based on Schott's approximation appear too conservative, as their type I error rates fall below the nominal level 0.05, see Table 1. Instead the type I error rate of the *Bootstrap* test ranges between 0.054 and 0.079. In this scenario, all the considered tests show a power close to 1, except *Schott est.*,

		(n_y, n_z)								
		(20,20)	(20,60)	(20,100)	(60,20)	(60,60)	(60,100)	(100,20)	(100,60)	(100,100)
Gaussian	Schott theo.	0.994	1.000	1.000	0.998	1.000	1.000	0.998	1.000	1.000
	Schott est.	0.963	0.985	0.996	0.994	0.997	0.991	0.991	1.000	0.996
	Bootstrap	0.987	1.000	1.000	0.996	1.000	1.000	0.991	1.000	1.000
Student's t (40 df)	Schott theo.	0.995	0.999	0.999	0.999	1.000	0.998	0.994	1.000	1.000
	Schott est.	0.964	0.976	0.997	0.994	0.992	0.989	0.994	1.000	0.995
	Bootstrap	0.985	0.998	0.998	0.997	1.000	1.000	0.996	1.000	1.000
Student's t (8 df)	Schott theo.	0.993	0.998	1.000	0.998	1.000	0.996	0.996	1.000	1.000
	Schott est.	0.955	0.981	0.995	0.988	0.994	0.986	0.990	1.000	0.995
	Bootstrap	0.976	0.999	0.998	0.989	0.999	1.000	0.988	0.999	1.000
Student's t (4 df)	Schott theo.	0.993	0.999	1.000	0.995	1.000	0.997	0.991	1.000	1.000
	Schott est.	0.969	0.985	0.992	0.991	0.992	0.988	0.985	0.999	0.994
	Bootstrap	0.977	0.990	0.992	0.986	0.996	0.997	0.982	0.998	1.000
Uniform	Schott theo.	0.999	1.000	0.999	0.998	1.000	0.997	0.998	1.000	1.000
	Schott est.	0.965	0.988	0.998	0.994	0.996	0.982	0.997	1.000	0.995
	Bootstrap	0.994	1.000	0.999	0.994	1.000	1.000	0.999	1.000	1.000

Table 2: Estimated power of the tests for varying sample sizes and sampling probability. Data were simulated according to the alternative hypothesis as described in scenario $S2$. The nominal value of tests is 5%.

which shows lower power at the smallest sample sizes.

5. Real-data example

The real-data example presented here concerns a respiratory microbiome measurement in two groups of patients. The data are provided in aggregate form, as the experiment from which they originate is still ongoing and full disclosure is not yet authorized. The first group (G1) consists of 9 patients with a severe disease who received no treatment, while the second group (G2) includes 7 patients undergoing antibiotic therapy.

To illustrate the application of the theory, we selected only the 7 most represented phyla, which account for the 99.97% of the total counts. In the excluded 14 phyla more than half the counts are zeros. In 2 of the selected phyla, the bacteria are completely eliminated by the antibiotic therapy in group G2, resulting in two columns of structural zeros. Furthermore, the drug affects all the phyla, leading to a considerable reduction in bacterial counts across the board in the second group. Table 5 displays the mean and

		(n_y, n_z)								
		(20,20)	(20,60)	(20,100)	(60,20)	(60,60)	(60,100)	(100,20)	(100,60)	(100,100)
Gaussian	Schott theo.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Schott est.	0.959	0.978	0.993	0.997	0.987	0.992	0.999	0.999	0.993
	Bootstrap	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
Student's t (40 df)	Schott theo.	1.000	1.000	1.000	1.000	1.000	0.996	1.000	1.000	1.000
	Schott est.	0.967	0.977	0.993	1.000	0.987	0.987	0.999	0.999	0.994
	Bootstrap	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Student's t (8 df)	Schott theo.	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000
	Schott est.	0.957	0.976	0.995	0.998	0.986	0.988	0.999	1.000	0.991
	Bootstrap	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Student's t (4 df)	Schott theo.	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000
	Schott est.	0.959	0.978	0.995	0.995	0.986	0.989	0.996	0.997	0.994
	Bootstrap	0.998	1.000	1.000	1.000	1.000	1.000	0.999	0.999	1.000
Uniform	Schott theo.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Schott est.	0.950	0.978	0.997	1.000	0.993	0.987	1.000	0.999	0.990
	Bootstrap	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 3: Estimated power of the tests for varying sample sizes and sampling probability. Data were simulated according to the alternative hypothesis as described in scenario $S3$. The nominal value of tests is 5%.

standard deviation of the counts for each of the 7 phyla considered in both groups.

The objective is to assess whether a common structure, as defined in the previous sections, can still be identified in the two groups despite the significant reduction in absolute counts. After computing the closure to 1 of each sample, we tested the null hypothesis in Eq. (11) with $K \in \{1, 2, 3, 4\}$. We used both, the modified Schott's test for normally distributed data described in Section 3.1 and the bootstrap procedure described in Section 3.2, with $n_{boot} = 1000$ bootstrap samples. When setting the significance level to 0.05, both tests always rejected the null hypothesis (the p-value is $P \leq 0.001$ for $K = 1, 3, 4$, while for $K = 2$ we have $P = 0.035$ with the modified Schott's test, and $P = 0.005$ with the bootstrap test). The results thus suggest that there is no common subspace between the two groups.

Phylum	First group (G1)	Second group (G2)
Ph1	2931.89 (2030.41)	1918.56 (2128.84)
Ph2	21797.33 (40020.97)	1201.67 (2656.75)
Ph3	1608.11 (4297.38)	138.89 (107.19)
Ph4	15583.11 (12291.77)	3762.56 (3599.72)
Ph5	6399.22 (8777.69)	482.44 (334.38)
Ph6	31.56 (33.035)	0 (0)
Ph7	65.11 (81.75)	0 (0)

Table 4: Mean (standard deviation) of the bacterial counts for the 7 selected phyla in the two patient groups.

6. Discussion

In this study, we introduced a statistical procedure to test for a common principal component subspace between two compositional datasets, one of which contains structural zeros. In details, we adapted Schott’s test (Schott, 1988) to deal with compositional datasets normally distributed on the simplex, while we proposed a bootstrap test for the non parametric case.

Our simulations show that the bootstrap test provides results comparable to the adapted Schott’s test when data are sampled from Gaussian distributions on the simplex. Conversely, in the general case, the bootstrap approach outperforms Schott’s test in the sense that it shows an higher power while keeping the type I error rate closer to the nominal level of the test.

We applied our approach to a real-data example concerning respiratory microbiome measurements from a group of patients undergoing antibiotic therapy and a control group who did not receive any treatment. In the former group two of the seven considered phyla are structural zeros. Both the adapted Schott’s test and the bootstrap test rejected the hypothesis of a common subspace between the two groups, suggesting that the antibiotic treatment has a substantial effect on the whole composition of the microbiome. Despite the limited sample size of our datasets, this example illustrates how our methods can be applied in a real-world setting. Future efforts will focus on testing our procedures across a broader range of applications.

In this paper we focused on the comparison between two datasets. Future effort will be devoted in testing for a principal subspace common to three or more datasets, each one possibly involving different sets of variables that are structural zeros.

Acknowledgments

The research of F.P. and S.S. was partially support by the project “DHEAL-COM - Digital Health Solutions in Community Medicine” (CUP: D33C22001980001) under the Innovative Health Ecosystem (PNC)—National Recovery and Resilience Plan (NRRP) program funded by the Italian Ministry of Health.

F.P. and F.R. are members of the INdAM-GNAMPa group, S.S. is member of the INdAM-GNCS group.

Appendix A. Null distribution of $\mathbf{T}$ when data are normally distributed on the simplex.

Appendix A.1. Auxiliary results

For the proof of Theorem 1 we make use of the following properties of the sample covariance matrices (cfr. Bishop et al., 2018).

Theorem 2. *Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be n i.i.d. P -dimensional random vectors following a multivariate normal distribution $\mathcal{N}_P(\boldsymbol{\mu}, \boldsymbol{\Omega})$. Denoted with $\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$ and with $\hat{\boldsymbol{\Omega}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$ the corresponding sample mean and sample covariance matrix it holds*

$$(n-1)\hat{\boldsymbol{\Omega}} \sim \mathcal{W}_P(\boldsymbol{\Omega}, n-1).$$

Theorem 3. *If $\mathbf{S} \sim \mathcal{W}_P(\boldsymbol{\Omega}, m)$ and $\mathbf{C} \in \mathbb{R}^{Q \times P}$ is a matrix with rank P , then $\mathbf{CSC}' \sim \mathcal{W}_P(\mathbf{C}\boldsymbol{\Omega}\mathbf{C}', m)$.*

Theorem 4. *Assume $\mathbf{S} \sim \mathcal{W}_P(\text{diag}(\delta_1, \dots, \delta_P), n)$ with $\delta_1 \geq \delta_2 \geq \dots \geq \delta_K > \delta_{K+1} \geq \dots \geq \delta_P$ and define $\mathbf{W} = \mathbf{S} - n \text{diag}(\delta_1, \dots, \delta_P)$. Then, up to second order terms in the elements of $n^{-1}\mathbf{W}$, it holds*

$$\sum_{i=1}^K \lambda_i(\mathbf{S}) \simeq n \left(\sum_{i=1}^K \delta_i + \sum_{i=1}^K \frac{w_{ii}}{n} + \sum_{i=1}^K \sum_{j=K+1}^{D-1} \frac{w_{ij}^2}{n^2(\delta_i - \delta_j)} \right)$$

where $\lambda_i(\mathbf{S})$ denotes the i -th greatest eigenvalue of $\mathbf{S}$.

Appendix A.2. Proof of Theorem 1

Proof. For the ease of exposition, henceforth we denote

$$\mathbf{U}^{\hookrightarrow} = \begin{bmatrix} \mathbf{U} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-Q-1)} & \mathbf{I}_Q \end{bmatrix},$$

and given a matrix $\mathbf{M}$ we denote with $\mathbf{m}_i$ its i -th column and with $\text{tr}(\mathbf{M})$ its trace.

We observe that if the null hypothesis $\mathbf{H}_0$ in Eq. (11) is true then there exist orthogonal matrices $\mathbf{R}_1 \in \mathbb{R}^{K \times K}$ and $\mathbf{R}_2 \in \mathbb{R}^{(D-K-1) \times (D-K-1)}$ such that $(\mathbf{u}_1^{\hookrightarrow}, \dots, \mathbf{u}_K^{\hookrightarrow}) = (\mathbf{v}_1, \dots, \mathbf{v}_K) \mathbf{R}_1$ and $(\mathbf{u}_{K+1}^{\hookrightarrow}, \dots, \mathbf{u}_{D-1}^{\hookrightarrow}) = (\mathbf{v}_{K+1}, \dots, \mathbf{v}_{D-1}) \mathbf{R}_2$. By exploiting these equalities into Eq. (18) we can rewrite the pooled covariance matrix as

$$\mathbf{\Omega}_{pool} = \mathbf{V} \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0}_{K,D-K-1} \\ \mathbf{0}_{D-K-1,K} & \mathbf{\Sigma}_2 \end{bmatrix} \mathbf{V}'$$

where $\mathbf{\Sigma}_1 \in \mathbb{R}^{K \times K}$ and $\mathbf{\Sigma}_2 \in \mathbb{R}^{(D-K-1) \times (D-K-1)}$ are symmetric matrices defined as

$$\mathbf{\Sigma}_1 = \mathbf{R}_1 \frac{n_y - 1}{n_y + n_z - 2} \text{diag}(\alpha_1, \dots, \alpha_K) \mathbf{R}_1' + \frac{n_z - 1}{n_y + n_z - 2} \text{diag}(\beta_1, \dots, \beta_K)$$

$$\mathbf{\Sigma}_2 = \mathbf{R}_2 \frac{n_y - 1}{n_y + n_z - 2} \text{diag}(\alpha_{K+1}, \dots, \alpha_{D-Q-1}, 0, \dots, 0) \mathbf{R}_2' + \frac{n_z - 1}{n_y + n_z - 2} \text{diag}(\beta_{K+1}, \dots, \beta_{D-1})$$

It follows that, under $\mathbf{H}_0$, the following two properties hold:

- (i) the K greatest eigenvalues of $\mathbf{\Omega}_{pool}$ correspond to the eigenvalues of $\mathbf{\Sigma}_1$, and hence

$$\sum_{i=1}^K \psi_i = \text{tr}(\mathbf{\Sigma}_1) = \frac{n_y - 1}{n_y + n_z - 2} \sum_{i=1}^K \alpha_i + \frac{n_z - 1}{n_y + n_z - 2} \sum_{i=1}^K \beta_i; \quad (\text{A.1})$$

- (ii) the normalized eigenvectors of $\mathbf{\Omega}_{pool}$ are of the form

$$\mathbf{K} = \mathbf{V} \begin{bmatrix} \mathbf{Q}_1 & \mathbf{0}_{K,D-K-1} \\ \mathbf{0}_{D-K-1,K} & \mathbf{Q}_2 \end{bmatrix},$$

being $\mathbf{Q}_1$ and $\mathbf{Q}_2$ the matrices of the normalized eigenvectors of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, respectively. As a consequence, both $\mathbf{V}^*$ and $\mathbf{U}^*$ are block diagonal matrices of the form

$$\mathbf{V}^* = \begin{bmatrix} \mathbf{V}_1^* & \mathbf{0}_{K,D-K-1} \\ \mathbf{0}_{D-K-1,K} & \mathbf{V}_2^* \end{bmatrix} \quad \text{and} \quad \mathbf{U}^* = \begin{bmatrix} \mathbf{U}_1^* & \mathbf{0}_{K,D-K-1} \\ \mathbf{0}_{D-K-1,K} & \mathbf{U}_2^* \end{bmatrix}.$$

Since by hypothesis $\mathbf{y}_1, \dots, \mathbf{y}_{n_y}$ and $\mathbf{z}_1, \dots, \mathbf{z}_{n_z}$ are realizations of random compositions normally distributed on the simplex, the corresponding pivot logratio coordinates are realization of Gaussian random vectors and hence Theorem 2 and Theorem 3 can be applied to the corresponding sample covariance matrices. This leads to the following results:

$$(n_y - 1)\mathbf{U}'\widehat{\boldsymbol{\Omega}}_{\mathbf{Y}}\mathbf{U} \sim \mathcal{W}_{D-Q-1}(\text{diag}(\alpha_1, \dots, \alpha_{D-Q-1}), n_y - 1) \quad (\text{A.2})$$

and

$$(n_z - 1)\mathbf{V}'\widehat{\boldsymbol{\Omega}}_{\mathbf{Z}}\mathbf{V} \sim \mathcal{W}_{D-1}(\text{diag}(\beta_1, \dots, \beta_{D-1}), n_z - 1) . \quad (\text{A.3})$$

In order to apply Theorem 4, we define the matrix of size $(D-1) \times (D-1)$

$$\begin{aligned} \mathbf{A}^{\hookrightarrow} &:= (n_y - 1) \left[(\mathbf{KU}^*)' \begin{bmatrix} \widehat{\boldsymbol{\Omega}}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-1)} & \end{bmatrix} (\mathbf{KU}^*) - \text{diag}(\alpha_1, \dots, \alpha_{D-Q-1}, 0, \dots, 0) \right] \\ &= (n_y - 1) \left[(\mathbf{U}^{\hookrightarrow})' \begin{bmatrix} \widehat{\boldsymbol{\Omega}}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-1)} & \end{bmatrix} \mathbf{U}^{\hookrightarrow} - \text{diag}(\alpha_1, \dots, \alpha_{D-Q-1}, 0, \dots, 0) \right] \\ &= \begin{bmatrix} \mathbf{A} & \mathbf{0}_{(D-Q-1),Q} \\ \mathbf{0}_{Q,(D-1)} & \end{bmatrix} \end{aligned} \quad (\text{A.4})$$

where

$$\mathbf{A} := (n_y - 1) \left[\mathbf{U}'\widehat{\boldsymbol{\Omega}}_{\mathbf{Y}}\mathbf{U} - \text{diag}(\alpha_1, \dots, \alpha_{D-Q-1}) \right] .$$

Similarly we define

$$\begin{aligned} \mathbf{B} &:= (n_z - 1) \left[(\mathbf{KV}^*)'\widehat{\boldsymbol{\Omega}}_{\mathbf{Z}}(\mathbf{KV}^*) - \text{diag}(\beta_1, \dots, \beta_{D-1}) \right] \\ &= (n_z - 1) \left[\mathbf{V}'\widehat{\boldsymbol{\Omega}}_{\mathbf{Z}}\mathbf{V} - \text{diag}(\beta_1, \dots, \beta_{D-1}) \right] . \end{aligned}$$

By applying Theorem 4 to the matrices in Eqs. (A.3) and (A.2) we obtain that, up to second order terms in the element of $(n_z - 1)^{-1}\mathbf{B}$ and $(n_y - 1)^{-1}\mathbf{A}$, respectively,

$$(n_z - 1) \sum_{i=1}^K \widehat{\beta}_i \simeq (n_z - 1) \sum_{i=1}^K \beta_i + \sum_{i=1}^K b_{ii} + \sum_{i=1}^K \sum_{j=K+1}^{D-1} \frac{b_{ij}^2}{(n_z - 1)(\beta_i - \beta_j)} \quad (\text{A.5})$$

and

$$\begin{aligned}
(n_y - 1) \sum_{i=1}^K \hat{\alpha}_i &\simeq (n_y - 1) \sum_{i=1}^K \alpha_i + \sum_{i=1}^K a_{ii} + \sum_{i=1}^K \sum_{j=K+1}^{D-Q-1} \frac{a_{ij}^2}{(n_y - 1)(\alpha_i - \alpha_j)} \\
&= (n_y - 1) \sum_{i=1}^K \alpha_i + \sum_{i=1}^K a_{ii}^{\hookrightarrow} + \sum_{i=1}^K \sum_{j=K+1}^{D-1} \frac{(a_{ij}^{\hookrightarrow})^2}{(n_y - 1)(\alpha_i - \alpha_j)}
\end{aligned} \tag{A.6}$$

where in the last equality we set $\alpha_j = 0$ for all $j = D - Q - 1, \dots, D - 1$, and we exploited Eq. (A.4) which guarantees that, for all $i = 1, \dots, K$,

$$a_{ij}^{\hookrightarrow} = \begin{cases} a_{ij} & \text{for all } j = i, \dots, D - Q - 1 \\ 0 & \text{for all } j = D - Q, \dots, D - 1 \end{cases} .$$

We further observe that

$$\begin{aligned}
\mathbf{U}^* \mathbf{A}^{\hookrightarrow} \mathbf{U}^{*'} + \mathbf{V}^* \mathbf{B} \mathbf{V}^{*'} &= \\
&= (n_y - 1) \mathbf{K}' \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1), Q} \\ \mathbf{0}_{Q, (D-1)} \end{bmatrix} \mathbf{K} - (n_y - 1) \mathbf{U}^* \text{diag}(\alpha_1, \dots, \alpha_{D-Q-1}, 0, \dots, 0) \mathbf{U}^{*'} \\
&\quad + (n_z - 1) \mathbf{K}' \hat{\Omega}_{\mathbf{Z}} \mathbf{K} - (n_z - 1) \mathbf{V}^* \text{diag}(\beta_1, \dots, \beta_{D-1}) \mathbf{V}^{*'} \\
&= \mathbf{K}' \left[(n_y - 1) \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1), Q} \\ \mathbf{0}_{Q, (D-1)} \end{bmatrix} + (n_z - 1) \hat{\Omega}_{\mathbf{Z}} \right] \mathbf{K} \\
&\quad - \mathbf{K}' \left[(n_y - 1) \begin{bmatrix} \Omega_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1), Q} \\ \mathbf{0}_{Q, (D-1)} \end{bmatrix} + (n_z - 1) \Omega_{\mathbf{Z}} \right] \mathbf{K} \\
&= \mathbf{K}' \left[(n_y - 1) \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1), Q} \\ \mathbf{0}_{Q, (D-1)} \end{bmatrix} + (n_z - 1) \hat{\Omega}_{\mathbf{Z}} \right] \mathbf{K} \\
&\quad - (n_y + n_z - 2) \mathbf{K}' \Omega_{\text{pool}} \mathbf{K} \\
&= \mathbf{K}' \left[(n_y - 1) \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1), Q} \\ \mathbf{0}_{Q, (D-1)} \end{bmatrix} + (n_z - 1) \hat{\Omega}_{\mathbf{Z}} \right] \mathbf{K} \\
&\quad - (n_y + n_z - 2) \text{diag}(\psi_1, \dots, \psi_{D-1}) .
\end{aligned}$$

Hence, following the same procedure described in Schott (1988), by exploiting the block diagonal form of $\mathbf{U}^*$ and $\mathbf{V}^*$, it is possible to obtain

$$\begin{aligned}
\sum_{i=1}^K \hat{\gamma}_i &= \sum_{i=1}^K \lambda_i \left(\mathbf{K}' \left[(n_y - 1) \begin{bmatrix} \hat{\Omega}_{\mathbf{Y}} & \mathbf{0}_{(D-Q-1), Q} \\ \mathbf{0}_{Q, (D-1)} \end{bmatrix} + (n_z - 1) \hat{\Omega}_{\mathbf{Z}} \right] \mathbf{K} \right) \\
&\simeq (n_y + n_z - 2) \sum_{i=1}^K \psi_i + \sum_{i=1}^K (a_{ii}^{\leftrightarrow} + b_{ii}) + \sum_{i=1}^K \sum_{j=K+1}^{D-1} \frac{(\mathbf{u}_i^* \mathbf{A}_{12}^{\leftrightarrow} \mathbf{u}_j^{*'} + \mathbf{v}_i^* \mathbf{B}_{12} \mathbf{v}_j^{*'})^2}{(n_y + n_z - 2)(\psi_i - \psi_j)}
\end{aligned} \tag{A.7}$$

where we denoted with $\mathbf{u}_1^*, \dots, \mathbf{u}_K^*$ and $\mathbf{v}_1^*, \dots, \mathbf{v}_K^* \in \mathbb{R}^K$ the rows of $\mathbf{U}_1^*$ and $\mathbf{V}_1^*$, with $\mathbf{u}_{K+1}^*, \dots, \mathbf{u}_{D-1}^*$ and $\mathbf{v}_{K+1}^*, \dots, \mathbf{v}_{D-1}^* \in \mathbb{R}^{D-K-1}$ the rows of $\mathbf{U}_2^*$ and $\mathbf{V}_2^*$, and with $\mathbf{A}_{12}^{\leftrightarrow}$ and $\mathbf{B}_{12}$ the matrices of size $K \times (D - K - 1)$ comprising the first K rows and the last $D - K - 1$ columns of $\mathbf{A}^{\leftrightarrow}$ and $\mathbf{B}$, respectively.

By replacing the approximation derived in Eqs. A.5, (A.6), and (A.7) into the definition of the test statistics T provided in Eq. (13) and by exploiting Eq. (A.1) we obtain that under $\mathbf{H}_0$ we can approximate

$$T \simeq \sum_{i=1}^K \sum_{j=K+1}^{D-1} \left[\frac{(a_{ij}^{\leftrightarrow})^2}{(n_y - 1)(\alpha_i - \alpha_j)} + \frac{b_{ij}^2}{(n_z - 1)(\beta_i - \beta_j)} - \frac{(\mathbf{u}_i^* \mathbf{A}_{12}^{\leftrightarrow} \mathbf{u}_j^{*'} + \mathbf{v}_i^* \mathbf{B}_{12} \mathbf{v}_j^{*'})^2}{(n_y + n_z - 2)(\psi_i - \psi_j)} \right]. \tag{A.8}$$

If we further assume $\Omega_{\mathbf{Y}}$ and $\Omega_{\mathbf{Z}}$ to be diagonal, it can be easily shown that: $\mathbf{K} = \mathbf{U}^* = \mathbf{V}^* = \mathbf{I}_{D-1}$, $\psi_i = \frac{(n_y-1)\alpha_i + (n_z-1)\beta_i}{n_y + n_z - 1}$ for all $i = 1, \dots, D - Q - 1$, and $\psi_i = \frac{(n_z-1)\beta_i}{n_y + n_z - 1}$ for all $i = D - Q, \dots, D - 1$. Hence the approximation of T becomes

$$T \simeq \sum_{i=1}^K \sum_{j=K+1}^{D-1} \frac{(n_y - 1)(n_z - 1) (a_{ij}^{\leftrightarrow}(\beta_i - \beta_j)/(n_y - 1) - b_{ij}(\alpha_i - \alpha_j)/(n_z - 1))^2}{(\alpha_i - \alpha_j)(\beta_i - \beta_j)[(n_y - 1)(\alpha_i - \alpha_j) + (n_z - 1)(\beta_i - \beta_j)]}. \tag{A.9}$$

It can be easily shown that the right-hand side of Eq. (A.9) is asymptotically distributed as a linear combination of independent chi-squared random variables. Therefore, as suggested by Schott (1988), also in the presence of structural zeros the test statistics T can be approximated as in Eq. (15) where μ_T and σ_T^2 are obtained by computing mean and variance of the right-hand side of Eq. (A.8). $\square$

References

- Aitchison, J., 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 44, 139–160. doi:10.1111/j.2517-6161.1982.tb01195.x.
- Aitchison, J., 1983. Principal component analysis of compositional data. *Biometrika* 70, 57–65. doi:10.1093/biomet/70.1.57.
- Aitchison, J., 1984. Reducing the dimensionality of compositional data sets. *Mathematical Geology* 16, 617–635. doi:10.1007/BF01029321.
- Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. *Journal of the Royal Statistical Society, Series C: Applied Statistics* 51, 375–392. doi:10.1111/1467-9876.00275.
- Alenazi, A., 2021. A review of compositional data analysis and recent advances. *Communications in Statistics - Theory and Methods* 52, 5535–5567. doi:10.1080/03610926.2021.2014890.
- Bishop, A.N., Del Moral, P., Niclas, A., 2018. An introduction to Wishart matrix moments. *Foundations and Trends in Machine Learning* 11, 97–218. doi:10.1561/22000000072.
- Cicchella, D., Ambrosino, M., Gramazio, A., Coraggio, F., Musto, M.A., Caputi, A., Avagliano, D., Albanese, S., 2022. Using multivariate compositional data analysis (CoDA) and clustering to establish geochemical backgrounds in stream sediments of an onshore oil deposits area. the Agri River basin (Italy) case study. *Journal of Geochemical Exploration* 238, 107012. doi:10.1016/j.gexplo.2022.107012.
- Cuesta-Albertos, J.A., Cuevas, A., Fraiman, R., 2009. On projection-based tests for directional and compositional data. *Statistics and Computing* 19, 367–380. doi:10.1007/s11222-008-9098-3.
- Egozcue, J.J., Pawlowsky-Glahn, V., 2019. Compositional data: the sample space and its structure. *TEST* 28, 599–638. doi:10.1007/s11749-019-00670-6.
- Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. *Mathematical geology* 35, 279–300. doi:10.1023/A:1023818214614.

- Filzmoser, P., Hron, K., Reimann, C., 2009. Principal component analysis for compositional data with outliers. *Environmetrics* 20, 621–632. doi:10.1002/env.966.
- Filzmoser, P., Hron, K., Templ, M., 2018. *Applied Compositional Data Analysis*. Springer, Cham.
- Fiori, A.M., Porro, F., 2023. A compositional analysis of systemic risk in european financial institutions. *Annals of Finance* 19, 325–354. doi:10.1007/s10436-023-00427.
- Fisher, N.I., 1993. *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge.
- Greenacre, M., 2018. *Compositional data analysis in practice*. CRC Press, Boca Raton, FL.
- Greenacre, M., 2024. The chiPower transformation: a valid alternative to logratio transformations in compositional data analysis. *Advances in Data Analysis and Classification* 18, 769–796. doi:10.1007/s11634-024-00600-x.
- Grifoll, M., Ortego, M., Egozcue, J.J., 2019. Compositional data techniques for the analysis of the container traffic share in a multi-port region. *European Transport Research Review* 11, 1–15. doi:10.1186/s12544-019-0350-z.
- Kim, K., Park, J., Jung, S., 2024. Principal component analysis for zero-inflated compositional data. *Computational Statistics and Data Analysis* 198, 107989. doi:10.1016/j.csda.2024.107989.
- Lee, A., 2010. Circular data. *WIREs Computational Statistics* 2, 477–486. doi:10.1002/wics.98.
- Lu, S., Wang, W., Guan, R., 2024. Kent feature embedding for classification of compositional data with zeros. *Statistics and Computing* 34, 34–69. doi:10.1007/s11222-024-10382-z.
- Lubbe, S., Filzmoser, P., Templ, M., 2021. Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems* 210, 104248. doi:10.1016/j.chemolab.2021.104248.

- Martin-Fernandez, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J., 2012. Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics and Data Analysis* 56, 2688–2704. doi:10.1016/j.csda.2012.02.012.
- Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2011. The Principle of Working on Coordinates. John Wiley & Sons, Ltd. chapter 3. pp. 29–42. doi:10.1002/9781119976462.ch3.
- Mert, M.C., Filzmoser, P., Hron, K., 2015. Sparse principal balances. *Statistical Modelling* 15, 159–174. doi:10.1177/1471082X1453552.
- Mezzadri, F., 2007. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society* 54, 592–604.
- Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and analysis of compositional data. John Wiley & Sons, Chichester.
- Pewsey, A., García-Portugués, E., 2021. Recent advances in directional statistics. *TEST* 30, 1–58. doi:10.1007/s11749-021-00759-x.
- Pewsey, A., Neuhäuser, M., Ruxton, G.D., 2013. Circular statistics in R. Oxford University Press, Oxford.
- Porro, F., Rapallo, F., Sommariva, S., 2024. Dimensionality reduction of compositional data with structural zeros: A case study, in: Scientific Meeting of the Italian Statistical Society, Springer. pp. 107–112.
- Rieser, C., Filzmoser, P., 2023. Extending compositional data analysis from a graph signal processing perspective. *Journal of Multivariate Analysis* 198, 105209. doi:10.1016/j.jmva.2023.105209.
- Scealy, J.L., De Caritat, P., Grunsky, E.C., Tsagris, M.T., Welsh, A.H., 2015. Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association* 110, 136–148. doi:10.1080/01621459.2014.990563.
- Schott, J.R., 1988. Common principal component subspaces in two groups. *Biometrika* 75, 229–236. doi:10.1093/biomet/75.2.229.

- Stewart, G.W., 1980. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis* 17, 403–409. doi:10.1137/0717034.
- Tsagris, M.T., Preston, S., Wood, A.T., 2011. A data-based power transformation for compositional data, in: *Proceedings of the 4th Compositional Data Analysis Workshop*, Girona, Spain.
- Tsilimigras, M.C., Fodor, A.A., 2016. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology* 26, 330–335. doi:10.1016/j.annepidem.2016.03.002.
- Wang, H., Shangquan, L., Guan, R., Billard, L., 2015. Principal component analysis for compositional data vectors. *Computational Statistics* 30, 1079–1096. doi:10.1007/s00180-015-0570-1.