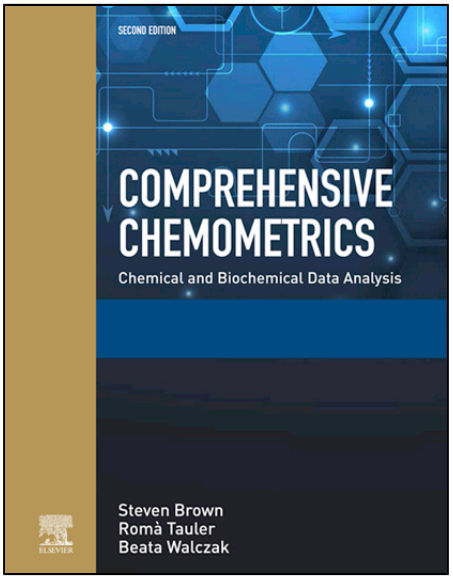

From Filzmoser, P.; Hron, K. Compositional Data Analysis in Chemometrics. In Comprehensive Chemometrics: Chemical and Biochemical Data Analysis; Brown, S., Tauler, R., Walczak, B., Eds., Elsevier, 2020; pp 641–662.
ISBN: 9780444641656

2.30 Compositional Data Analysis in Chemometrics

Peter Filzmoser, Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria
Karel Hron, Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Olomouc, Czech Republic

2.30.1 Introduction

Various attempts have been made in the literature for a mathematical definition of the term *compositional data*.[1,2] Here we will not start with a definition of this term, but rather provide a simple example of a *composition*. Think about a recipe for a cocktail. The recipe will include D different ingredients that need to be mixed together. In compositional data analysis (abbreviation "CoDA") one talks about D compositional parts. Further, the amount of each ingredient is defined in the recipe. This amount, however, depends on the overall quantity, related to the number of persons for which the cocktail is prepared. For twice as many people one would have to multiply each amount by 2. So, the taste of the cocktail is not depending on the amounts, but rather on the relative contributions of the D ingredients. If we want to modify the taste (with the given ingredients), then the ratio of one specific ingredient to another one needs to be changed. Overall, one could think in terms of all pairwise ratios between the ingredients to specify the taste.

More formally, consider an observation $\mathbf{x}_i = (x_{i1},\ldots, x_{iD})'$ with D compositional parts, and in total n observations are available, thus $i = 1,\ldots, n$. All together, there are $D(D-1)$ different pairwise ratios (by excluding the trivial ones formed by ratios of the parts to themselves):

$$\frac{x_{i1}}{x_{i2}}, \frac{x_{i1}}{x_{i3}}, \ldots, \frac{x_{i1}}{x_{iD}}, \frac{x_{i2}}{x_{i1}}, \frac{x_{i2}}{x_{i3}}, \ldots, \frac{x_{i,D-1}}{x_{iD}}.$$

It is not very elegant to consider the ratios together with their reverse values, for example, both x_{i1}/x_{i2} and x_{i2}/x_{i1}. Rather, it would be better to have a simple relation between them. Moreover, ratios are somehow asymmetric: while for a dominance of x_{i1} over x_{i2} through the ratio x_{i1}/x_{i2} the whole interval $(1, \infty)$ is reserved, in case of a dominance of x_{i2} over x_{i1} just $(0,1)$ must suffice. A way to

Comprehensive Chemometrics, Second Edition, 2020, 641–662

symmetrize the ratios is not new—just take a logarithm! Now, the *log-ratio* $\ln(x_{i1}/x_{i2}) = -\ln(x_{i2}/x_{i1})$, and it is therefore fully sufficient to consider $D(D-1)/2$ terms

$$\ln\left(\frac{x_{i1}}{x_{i2}}\right), \ln\left(\frac{x_{i1}}{x_{i3}}\right), \ldots, \ln\left(\frac{x_{i1}}{x_{iD}}\right), \ln\left(\frac{x_{i2}}{x_{i3}}\right), \ldots, \ln\left(\frac{x_{i,D-1}}{x_{iD}}\right) \tag{1}$$

to capture the elemental information of the data. Obviously, there is much redundant information in the above set of log-ratios. It is not difficult to show that any log-ratio can be expressed as linear combination of at most $D-1$ other log-ratios.[3] For example, for $D = 4$,

$$\ln\left(\frac{x_{i1}}{x_{i2}}\right) = \ln\left(\frac{x_{i1}}{x_{i4}}\right) - \ln\left(\frac{x_{i3}}{x_{i4}}\right) - \ln\left(\frac{x_{i2}}{x_{i3}}\right), \tag{2}$$

but also

$$\ln\left(\frac{x_{i1}}{x_{i2}}\right) = \ln\left(\frac{x_{i1}}{x_{i3}}\right) - \ln\left(\frac{x_{i2}}{x_{i3}}\right) \tag{3}$$

Of course, also in the latter case, the third log-ratio with a zero coefficient could be formally added. This implies that the space spanned by log-ratios has dimensionality $D-1$.

Coming back to the cocktail example, one can see that log-ratios are *scale invariant*, because they are not depending on the absolute amount of all involved ingredients. For example, preparing twice as much of the amount refers to a multiplication of each part by the factor 2, but this factor would not alter the log-ratio. A methodology based on log-ratios will thus produce the same results irrespective of the total (amount), or of the scale in which the data are expressed. This is referred to as the log-*ratio methodology*, and it will be introduced in detail in the next sections. Data that obey the scale invariance property are commonly known as *compositional data*,[4] therefore this method is referred to as the log-ratio methodology for compositional data.

From a more practical point of view, one could say that the relevant information contained in compositional data is in the relations between the compositional parts. For the cocktail example we refer here to the taste as the relevant issue, where the total amount of all ingredients together is not essential. However, the total amounts might still be important if somebody needs to buy the ingredients. In other words, the way how to analyze the data depends on the purpose of the analysis. In chemistry or chemometrics one is typically interested in identifying underlying processes generating the specific multivariate data structure. Moreover, spectral measurements are often re-scaled to sum up to 1% or 100%, because it is not the peak of the spectra itself which provides the relevant information, but rather the ratios. With the log-ratio methodology, a normalization to a constant sum 1 is irrelevant because of the scale invariance principle.

It was argued previously that $D-1$ log-ratios build the elemental information for the log-ratio methodology. However, it is of crucial importance how this elemental information contained in the log-ratios is processed. On the way of finding a reasonable approach, some immediate questions arise:

1. Can any $D-1$ log-ratios be used to represent compositional data with D components?
2. Could any interpretation in terms of the original parts be reached within the log-ratio methodology?
3. How to apply and interpret my favorite method using this approach?

The following sections will provide answers to these questions, but also to many others that are relevant to pose in case of compositional data.

2.30.1.1 The Concept of Compositional Data Analysis

In the previous section, the normalization of spectral measurements to 100 (in case of percentages) or 1 (for proportions) was mentioned. In several scientific communities, compositional data are still considered as *constrained data*, i.e. as observations that induce a constant sum constraint. The cocktail example clearly shows that this must not necessarily be the case. With the presented concept of compositional data, any sum of parts stands just for a particular representation of information that is contained in the ratios between the components. Formally, a representation of a composition $\mathbf{x}_i$, for any $i \in \{1, \ldots, n\}$, with an arbitrary but fixed constant sum κ can be done using the closure operation,

$$C_\kappa(\mathbf{x}_i) = \left(\frac{\kappa x_{i1}}{\sum_{j=1}^{D} x_{ij}}, \ldots, \frac{\kappa x_{iD}}{\sum_{j=1}^{D} x_{ij}}\right)'. \tag{4}$$

Accordingly, the sample space of compositional data (the formal space from where we pick up the samples) is formed by classes of proportional vectors from which any particular representation can be derived,

$$C^D = \left\{\mathbf{x} = (x_1, \ldots, x_D)' \in \mathbb{R}_+^D | \forall \kappa > 0 \ \exists! \ \lambda > 0 \ : \mathbf{x} = \lambda C_\kappa(\mathbf{x})\right\} \tag{5}$$

Because the sample space of constrained data is traditionally denoted as D-part simplex,

$$S^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)' \in \mathbb{R}_+^D | \sum_{j=1}^{D} x_j = \kappa \right\}, \tag{6}$$

definition (5) may refer to a *(D-part) extended simplex*. It is worth mentioning that both standard multivariate observations, where the absolute scale matters, and compositional data are picked up from the same positive real space $\mathbb{R}_+^D$. However, in the latter case this space is decomposed into the above classes of proportional vectors, leading to the extended simplex.

Another important point for the concept of compositional data analysis is the relative scale property. Suppose that a cocktail is composed from lemon juice, ananas juice and a spirit. For all these three ingredients, changes in the small concentrations are more important for the final taste of the drink. It makes a difference, whether the lemon juice is doubled from 5% to 10%—the cocktail may become too sour. On the other hand, having 50% or 55% of lemon juice does not play a role any more—in both cases the cocktail will be undrinkable! The difference between 5% and 10%, or 50% and 55% is the same (5%), but the ratio between them matters. While 10% is twice as much as 5%, 55% is just a 1.1 multiple of 50%. In other words, ratios play a crucial role not just *within* compositions, but also *between* them, and any reasonable metric should honor this feature.

Similarly as scale invariance, also the presence of relative scale may help to reveal whether the researcher is faced up to compositional data, or not. One particular example of constrained data is the RGB color model referring to an additive color model in which red, green and blue light are added together in various ways to reproduce a broad array of colors. In order to get a desired color, a value on the scale between 0 and 255 is assigned to the red, green and blue channels, while keeping the sum constraint 255 of the three components. However, here one can hardly say that smaller values assigned to individual color channels are more important than bigger ones—all colors resulting as a mixture of red, green and blue are of the same "importance". In this case, any approach that is able to deal with the specific properties of constrained data (see, e.g., ref. [5]) seems to be more appropriate.

Note that neither scale invariance nor relative scale of compositions as undetachable features of compositional data were explicitly mentioned in the seminal book on the log-ratio methodology by Aitchison[1] and are promoted only in monographs following recent trends in the field.[4,6] However, both these features inherently form the core of the concept of compositional data analysis.

2.30.1.2 Methodological Principles

The concept of compositional data analysis needs to be formalized in order to proceed with relevant (not exclusively) statistical analyses. Inspired by the above thoughts, Egozcue[2] introduced *principles of compositional data analysis* that may be listed as follows:

Scale invariance: The information in a composition does not depend on the particular units in which the composition is expressed. Proportional positive vectors represent the same composition. Any sensible characteristic of a composition should be invariant under a change of scale. This principle thus corresponds to the fact that a multiplication of a compositional vector by an arbitrary positive number does not alter the ratios between the compositional parts.

Permutation invariance: Permutation of parts of a composition does not alter the information conveyed by the compositional vector, similarly as in standard multivariate statistics.

Subcompositional coherence: Information conveyed by a composition of D parts should not be in contradiction with that coming from a subcomposition (i.e., a subvector of the original compositional vector) containing d parts, $d < D$. This principle can be formulated more precisely as.

- *Subcompositional dominance*: If $\Delta_p(\mathbf{x}, \mathbf{y})$ is any distance between compositions of p parts, then $\Delta_D(\mathbf{x}, \mathbf{y}) \geq \Delta_d(\mathbf{x}_d, \mathbf{y}_d)$, where $\mathbf{x}$, $\mathbf{y}$ are compositions with D parts and $\mathbf{x}_d$, $\mathbf{y}_d$ are subcompositions of the previous ones with d parts, $d < D$.
- *Ratio preserving*: Any relevant characteristic expressed as a function of the parts of a composition is exclusively a function of the ratios of its parts. In a subcomposition, these characteristics depend only on the ratios of the selected parts and not on the discarded parts of the parent composition. Scale invariance applies to the subcomposition.

While the principles of permutation invariance and subcompositional dominance should be fulfilled by *any* reasonable statistical analysis, aware of the corresponding geometrical consequences,[7] scale invariance is a specific principle resulting directly from the definition of compositional data.

The subcompositional coherence principle deserves more thorough explanations, because its definition in terms of *information conveyed by a composition of D parts should not be in contradiction with that coming from a subcomposition…* is indeed quite vague. The point is that compositional data analysis should behave like standard multivariate data analysis when just a subset of variables (parts) is considered. This is nicely represented by the subcompositional dominance principle, where the full objects (composition and its subcomposition, respectively) are taken. On the other hand, the subcompositional coherence principle might become misleading when referring to the original compositional parts. Due to scale invariance, it is clear that any part cannot stay alone, its value is determined by ratios with the other components. For example, the relative amount of the lemon juice component in the above cocktail is determined by its ratios with pineapple juice and spirit components, respectively. Accordingly, when one or more parts are omitted, and thus the *whole* to which the composition refers is changed, also the relative dominance of the part of interest within the new composition is necessarily changed. Therefore, in such cases one cannot refer any more to subcompositional (in)coherence, but rather to changing the set of elemental log-ratios (1) which are used to draw any statement (necessarily in the relative sense) about the original compositional part.

Ratio preserving implies also another important consequence: it is not possible to amalgamate (sum up, aggregate) any two parts of a composition without violating the scale invariance principle. In other words, when the amalgamation is done, it is neither possible to go back to the original parts, nor to change the order of the amalgamation and closure operation in the statistical analysis of compositions. Therefore, in line with the definition of compositional data, instead of amalgamating parts, one should focus on *amalgamation of* log-*ratios*. This concept is further developed in section ***Coordinate representations***.

2.30.1.3 Literature and Software

Since the seminal book on log-ratio analysis of compositional data was published,[1] a number of scientific papers was devoted to this topic reflecting a huge step forward that the methodology did in the last more than 30 years. Moreover, also several monographs were produced that aim to cover important aspects of compositional data analysis from different perspectives. Among them, Pawlowsky-Glahn et al.[4] can be currently considered as the reference book in the field. It contains the state-of-the-art of the log-ratio methodology and provides also guidelines for statistical modeling with compositional data. The book builds on previous developments collected, e.g., in contributed books by Buccianti et al.[8] and Pawlowsky-Glahn and Buccianti.[9] The latter one contains also contributions from various applications, though the majority of them coming from geosciences which formed the original source of motivation for the development of the log-ratio analysis. There are also two recent books focusing on practical aspects of compositional data analysis. While Greenacre[10] reflects rather personal views on recent developments in the log-ratio methodology and tries to bring it back to the roots as proposed by Aitchison,[1] Filzmoser et al.[6] aim to provide a broad and concise view on various aspects of compositional data processing, illustrated by examples from various applications (including chemometrics); similarly as in Greenacre,[10] all the proposed methods are accompanied by worked-out examples in the statistical software environment R.[11]

Developments in the log-ratio methodology are covered by packages in various software tools. Among them, contributed packages in R can be considered as those reflecting best the current developments. Some of them are listed below:

compositions: The oldest R package on compositional data analysis, including geometrical operations, descriptive exploratory analysis tools, multivariate statistical methods, and functionality for dealing with zeros and missing values. The package is thoroughly described in van den Boogaart and Tolosana-Delgado.[12]

robCompositions: This package includes classical and robust methods for dealing with compositional data, like outlier detection, principal component analysis and discriminant analysis, regression with compositional predictors, focusing on methods where some peculiarities can be expected when applying them to compositional data.[13] The package contains also algorithms for imputation and rounded zero replacement, and various log-ratio coordinate representations of compositional data that are considered in Filzmoser et al.[6] All examples presented in this article are computed with functions provided by this package.

zCompositions: The package serves for preprocessing issues. It contains functions for the treatment of zeros, left-censored and missing values in compositional data sets.[14]

easyCODA: This package covers univariate and multivariate methods for compositional data analysis, based on log-ratios, and it was developed to accompany the book Greenacre.[10]

There are also some more specific packages on compositional data analysis like **coda.base** (basic operations with compositional data including C++ implementations), **complmrob** (robust linear regression with compositional covariates according to Hron et al.[15]), and **propr** (identifying proportionally abundant features using compositional data analysis[16]).

For those who are not familiar with R, there is a freeware package CoDaPack that can be downloaded from the web site http://ima.udg.edu/CoDaPack. This point-and-click user interface relies on the Java Virtual Machine. It includes the basic mathematical operations with compositional data, log-ratio coordinate systems (see section ***Coordinate representations***), descriptive statistics and visualization tools.[17]

2.30.2 Coordinate Representations

According to Eq. (5), the sample space of compositional data is the set of equivalence classes of proportional vectors. The geometrical structure of compositions is called the *Aitchison geometry*,[4] and the goal here is to move the compositions from this geometry to the real Euclidean space, in which standard statistical tools can be used. This process is often referred to as *transformation*, and there are different kinds of such transformations available. Here we will rather refer to representing the compositions in coordinates, related to the real Euclidean space.

2.30.2.1 Geometrical Aspects

It is possible to define a vector space structure of the sample space C^D, defined through basic operations such as addition or multiplication. Using a specific naming convention and symbols for the notation of these operations, one can define:

- The **perturbation** of two compositions $\mathbf{x}$ and $\mathbf{y}$ from the sample space C^D as a composition

$$\mathbf{x} \oplus \mathbf{y} = (x_1 y_1, x_2 y_2, \ldots, x_D y_D)'. \tag{7}$$

- The power transformation (**powering**) of a composition $\mathbf{x} \in C^D$ by a constant $\alpha \in \mathbb{R}$ as

$$\alpha \odot \mathbf{x} = \left(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right)'. \tag{8}$$

These operations are sufficient to obtain a vector space.[4] The application of perturbation and powering allows to define the perturbation difference

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus [(-1) \odot \mathbf{y}] = (x_1/y_1, x_2/y_2, \ldots, x_D/y_D)'.$$

The perturbation difference between the same composition results in

$$\mathbf{x} \ominus \mathbf{x} = (x_1/x_1, x_2/x_2, \ldots, x_D/x_D)' = (1, 1, \ldots, 1)' = \mathbf{n},$$

which is called the *neutral element*.

With additional definitions of inner product, norm and distance in the Aitchison sense, one obtains a Euclidean vector space structure, in the literature denoted by the *Aitchison geometry*:

- The **Aitchison inner product**, defined for two compositions $\mathbf{x} = (x_1, \ldots, x_D)'$ and $\mathbf{y} = (y_1, \ldots, y_D)'$ from C^D as

$$\langle \mathbf{x}, y \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}. \tag{9}$$

- The **Aitchison norm** of a composition $\mathbf{x} = (x_1, \ldots, x_D)' \in C^D$ is defined by the inner product of $\mathbf{x}$ with itself,

$$\|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, x \rangle_A} = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j}\right)^2}. \tag{10}$$

- The Aitchison distance between x and y G C^D is defined as

$$d_A = (\mathbf{x}, y) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\right)^2}. \tag{11}$$

Note that all these definitions are based on log-ratios between the compositional parts, and thus one refers to the log-ratio methodology for compositional data analysis. As noted earlier, the goal is now to move the compositions from the Aitchison geometry to the real Euclidean space.

2.30.2.2 alr Coordinates

The abbreviation "alr" stands for "additive log-ratio," and this type of coordinates is mentioned rather for historical reasons.

Given a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ from C^D. Then alr coordinates $\mathbf{x}^{(j)}$ are represented in $\mathbb{R}^{D-1}$ and defined as

$$\mathbf{x}^{(j)} = \mathrm{alr}_j(\mathbf{x}) = \left(x_1^{(j)}, \ldots, x_{D-1}^{(j)}\right)' = \left(\ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j}\right)'. \tag{12}$$

The index $j \in \{1, \ldots, D\}$ refers to the variable in the denominator (ratioing variable), and the choice usually depends on the context. Clearly, alr coordinates will in general be different for another choice of the ratioing variable. A further limitation is that alr coordinates are not orthogonal to each other.[4]

2.30.2.3 clr Coefficients

"clr" stands for "centered log-ratio," and clr coefficients express a composition $\mathbf{x} \in C^D$ by a vector $\mathbf{y} \in \mathbb{R}^D$, defined as

$$\mathbf{y} = \mathrm{clr}(\mathbf{x}) = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{k=1}^{D} x_k}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{k=1}^{D} x_k}}\right)'. \tag{13}$$

The denominator in Eq. (13) is called the *geometric mean*. From a practical point of view, clr avoids the subjectivity of alr with respect to the choice of the denominator, and one also obtains D rather than only D–1 components. However, the sum of all clr coefficients is zero, and therefore one ends up with constrained data. This feature is emphasized by the terminology clr coefficients instead of clr coordinates. In fact, $\mathbf{y}$ represents coefficients with respect to a generating system, instead of a basis.[4]

As alr coordinates, the clr coefficients represent a one-to-one mapping, and the original parts—up to a scaling factor—are obtained as

$$x_j = \exp\left(y_j\right) \qquad \text{for} j = 1, \ldots, D. \tag{14}$$

The clr coefficients represent an *isometry*: For two compositions $\mathbf{x}_1$ and $\mathbf{x}_2 \in C^D$ it holds that.[18]

- $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A = \langle \mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2) \rangle$, $\|\mathbf{x}_1\|_A = \|\mathrm{clr}(\mathbf{x}_1)\|$;
- $d_A(\mathbf{x}_1, \mathbf{x}_2) = d(\mathrm{clr}(\mathbf{x}_1), \mathrm{clr}(\mathbf{x}_2))$.

Thus, the Aitchison inner product between the two compositions is the same as the usual inner product of the clr-quantities, and the Aitchison distance between two compositions is the Euclidean distance between the corresponding clr coefficients. In other words, all metric concepts in C^D are maintained when considering the clr coefficients. One also obtains linearity of the mapping:

$$\mathrm{clr}(\mathbf{x}_1 \oplus \mathbf{x}_2) = \mathrm{clr}(\mathbf{x}_1) + \mathrm{clr}(\mathbf{x}_2), \quad \mathrm{clr}(c \odot \mathbf{x}_1) = c \bullet \mathrm{clr}(\mathbf{x}_1)$$

for any $c \in \mathbb{R}$.

As mentioned above, from the perspective of the Aitchison geometry, alr variables form oblique coordinates (thus beeing not isometric) and clr variables are coefficients with respect to a generating system (isometric, but with a singular covariance matrix). Although both of them might be useful in specific cases, one should be aware of difficulties in these systems of variables. This is clear also when considering log-ratios as primary source of information in compositional data. Out of the set of simple log-ratios (1) where (up to sign) all possible combinations are contained, just D–1 are covered by alr coordinates. Each clr coefficient aggregates (up to sign) all log-ratios with x_j, $j = 1,\ldots, D$, but some log-ratios are contained in more coefficients simultaneously, an intuitive source of singularity of their covariance matrix. Thus, by considering question 1 posed at the end of ***Introduction*** section neither alr coordinates nor clr coefficients can be considered as fully satisfactory from the compositional perspective.

2.30.2.4 ilr and Pivot Coordinates

"ilr" stands for "isometric log-ratio," and this refers to a family of coordinates building an orthonormal basis in the D–1-dimensional hyperplane formed by clr coefficients.[18] One thus avoids the constraint of clr coefficients, and the resulting ilr coordinates are represented in $\mathbb{R}^{D-1}$. There are infinitely many possibilities to define such an orthonormal basis system, and one particular choice is called *pivot coordinates*, defined for a composition $\mathbf{x} \in C^D$ as

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$$

with

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}} \quad \text{for } j = 1, \ldots, D-1 \tag{15}$$

ref.[19]. With this choice, the part x_1 is only contained in the coordinate z_1, but in none of the remaining coordinates. Pivot coordinates therefore allow to separate one compositional part into one coordinate. This can be made even more explicit:

$$z_1 = \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\sqrt[D-1]{\prod_{k=2}^{D} x_k}} = \sqrt{\frac{1}{D(D-1)}} \left(\ln \frac{x_1}{x_2} + \ln \frac{x_1}{x_3} + \ldots + \ln \frac{x_1}{x_D} \right)$$

Coordinate z_1 summarizes all relative information about x_1 in terms of "averaged" log-ratios, and one can thus interpret this coordinate as dominance of x_1 with respect to the other parts "on average", if the values for z_1 are positive (and similarly in the other cases).

The above definition of pivot coordinates is particularly appropriate if the main interest is in the interpretation of z_1 in terms of x_1. Clearly, if the interest is in the interpretation of another part, say x_l, for $l \in \{2, \ldots, D\}$, then this part can be rearranged to the first position to obtain a composition $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$. Applying Eq. (15) to this composition leads to a first coordinate which isolates x_l[6] and, formally, to (generalized) pivot coordinates $z_1^{(l)}, \ldots, z_{D-1}^{(l)}$ for $l \in \{1, \ldots, D\}$. Different pivot coordinate systems (like any orthonormal coordinates) are rotations of each other that enables their flexible use for multivariate statistical processing.

Note that the clr coefficients and the first pivot coordinates are proportional to each other,

$$z_1^{(l)} = \sqrt{\frac{D}{D-1}} y_l, l = 1, \ldots, D \tag{16}$$

ref. [6], and thus one would obtain the same interpretation also for γ_l. However, x_l is still contained via the geometric mean in the other clr coefficients, and not isolated in one coordinate, as it is the case with $z_1^{(l)}$.

The *isometry* property is also valid for ilr coordinates[18]:

For two compositions $\mathbf{x}_1$ and $\mathbf{x}_2 \in C^D$ and $c \in \mathbb{R}$ it holds that

- $\mathrm{ilr}(\mathbf{x}_1 \odot \mathbf{x}_2) = \mathrm{ilr}(\mathbf{x}_1) + \mathrm{ilr}(\mathbf{x}_2), \quad \mathrm{ilr}(c \odot \mathbf{x}_1) = c \cdot \mathrm{ilr}(\mathbf{x}_1)$;
- $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_A = \langle \mathrm{ilr}(\mathbf{x}_1), \mathrm{ilr}(\mathbf{x}_2) \rangle, \quad \|\mathbf{x}_1\|_A = \|\mathrm{ilr}(\mathbf{x}_1)\|$;
- $d_A(\mathbf{x}_1, \mathbf{x}_2) = d(\mathrm{ilr}(\mathbf{x}_1), \mathrm{ilr}(\mathbf{x}_2))$.

Furthermore, also pivot coordinates represent a one-to-one mapping. The original parts can be (up to a scaling factor) uniquely reproduced from ilr co-ordinates,

$$\begin{aligned} x_1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1\right), \\ x_j &= \exp\left(-\sum_{k=1}^{j-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} z_k + \frac{\sqrt{D-j}}{\sqrt{D-j+1}} z_j\right), j = 2, \ldots, D-1, \\ x_D &= \exp\left(-\sum_{k=1}^{D-1} \frac{1}{\sqrt{(D-k+1)(D-k)}} z_k\right) \end{aligned} \tag{17}$$

From Eq. (15) it can be seen that any of the pairwise log-ratios (1) is aggregated uniquely in one of the pivot coordinates $z_1, \ldots, z_{D-1}$. This is another intuitive reason why pivot coordinates represent an appropriate choice of log-ratio coordinates also with respect to both questions 1 and 2 from the ***Introduction*** section.

2.30.2.5 Further Coordinates

As discussed in the previous section, pivot coordinates build an orthonormal basis in the hyperplane spanned by the clr coefficients, and they lead to an interpretation of one particular part, which is isolated in the first coordinate. The remaining coordinates are usually not easily interpretable. There are, however, other approaches for constructing an orthonormal basis, with different goals for the interpretation. We mention here two possibilities:

Symmetric pivot coordinates: The goal is to isolate two compositional parts into two coordinates, and the remaining (orthonormal) coordinates should not contain any information about these parts. In addition, the two specific coordinates treat the two parts symmetrically, see Kynčlová et al.[20] for details. Thus, the two coordinates contain all the relative information about the two parts of interest, and this information is represented in the standard Euclidean geometry. A scatterplots of the coordinates could then be of interest, revealing "appropriately" (with our Euclidean eyes) the relationships between the observations concerning these parts. Also correlation analysis can be carried out with these coordinates, and statistical hypothesis tests on uncorrelatedness can be performed. From a practical perspective it is interesting to note that correlations between symmetric pivot coordinates and between clr coefficients get more similar the higher the number of compositional parts is, see Kynčlová et al.[20]

Balances: The goal is to construct orthonormal coordinates, which represent non-overlapping groups of compositional parts. The resulting coordinates are called *balances*, since they refer to the balance between these groups.[21] The balances are constructed sequentially, and since they represent non-overlapping groups, the construction procedure is called *sequential binary partitioning*.[21] The results of this process can be visualized in the so-called CoDa dendrogram.[22] For balances, previous expert knowledge about grouping of parts is needed what is, however, rarely available with high-dimensional data. A data-driven counterpart to balances is called principal balances,[23] defined as a sequence of balances which successively maximize the explained variance in a data set. For high-dimensional problems, a sparse version of principal balances has been introduced, which limits the number of parts involved in the first few principal balances.[24]

2.30.3 Data Exploration

While it was not so difficult to find an appropriate answer for both questions 1 and 2 raised at the beginning of ***Introduction*** section by a proper choice of log-ratio coordinates, there is no simple answer to question 3 due to the complexity of the data structure which compositional data represent, and a wide range of problems the analytists are faced to. Although in principle it is possible to analyze compositional data using any popular multivariate method, when compositions are represented in appropriate log-ratio coordinates, one needs to take into account that any interpretation of results needs to be done in terms of (log-)ratios, even when linking the coordinates to the original parts like in case of pivot coordinates. This essential feature affects all levels of data processing, from exploratory data analysis and visualization to advanced tools for processing (high-dimensional) data, e.g. originating from chemometrics. The frequent appearance of high-dimensional observations in this field poses further questions that stimulate the development of the compositional methodology as such.

2.30.3.1 Statistical Data Summaries

Since all the relevant information in compositional data is contained in log-ratios, this should be reflected also by basic data summaries like the mean and the scatter estimate. In case of the mean, instead of the standard arithmetic mean its geometrical counterpart is preferred. The latter one enables to keep ratios under the requirement of scale invariance, i.e. with arbitrary representations of the input compositions. Indeed, let us come back again to the cocktail example, and consider the cocktail composed from lemon juice, ananas juice and a spirit again. Assume that the aim is to compute mean percentages of these ingredients from two samples of such cocktail, prepared by two barkeepers. Their cocktails differ not just in the relative composition of the cocktail, but also in the total volume (in ml). Accordingly, the first cocktail results in $\mathbf{x}_1 = (5, 150, 20)'$ and the second one in $\mathbf{x}_2 = (20, 180, 40)'$. The arithmetic mean computed from the original data results in $\bar{\mathbf{x}}^* = (12.5, 165, 30)'$, which gives in percentages $\bar{\mathbf{x}} = (6.0, 79.5, 14.5)'$. When the percentage representations of both samples would be taken, the "mean cocktail" composition would result in $\widetilde{\bar{\mathbf{x}}} = (5.6, 80.4, 14.0)'$. Obviously, this is a kind of inconsistent result, even if in the latter case the constant sum of the input data (100) is kept also for the resulting mean. The point is that the ratios between the components in both cases are different. On the other hand, the component-wise geometric mean does not depend on any particular representation of the input compositions and results (in its percentage representation) in $\mathbf{g} = (4.9, 81.1, 14.0)'$. It is also visible that the geometric mean (the so-called *center* in the compositional context) is more sensitive to the first component, where the relative scale matters most, compared to its arithmetic counterparts. Consider an $n \times D$ compositional data matrix $\mathbf{X} = (x_{ij})$ with the samples $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})'$ in its rows, $i = 1, \ldots, n$. Then the compositional center of $\mathbf{X}$ is defined as

$$\mathbf{g} = (g_1, \ldots, g_D)', \tag{18}$$

where $g_j = \left(\prod_{i=1}^{n} x_{ij}\right)^{1/n}$ is the geometric mean, for $j = 1, \ldots, D$. The center is also consistent with the Aitchison geometry and has analogous theoretical properties as the arithmetic mean in case of standard multivariate data. Importantly, the center corresponds to the arithmetic mean in any log-ratio coordinates. For example, for pivot coordinates (15) applied to the compositional matrix $\mathbf{X}$ it would be possible to get with the first coordinate a mean dominance of part x_1 (or any other component by permutation of the input composition) with respect to the averaged contributions of the other parts.

Also the usual scatter measures, variance and covariance, would not work with compositional data. For example, the empirical variance is defined as *arithmetic mean* of squared differences of the observations to the arithmetic mean. The same holds in a generalized form also for the covariances. Since the arithmetic mean turned out to be inappropriate with compositional data, an alternative scatter measure consistent with the Aitchison geometry is needed. It was found in a *variation matrix*,[1] which is formed by the variances of all pairwise log-ratios. In other words, the variation matrix does not attempt to work with the original parts like the compositional center, it turns directly to the elemental information in compositional data. Specifically, for the compositional data matrix $\mathbf{X} = (x_{ij})$, the variation matrix is defined as

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \ldots & t_{1D} \\ t_{21} & t_{22} & \ldots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \ldots & t_{DD} \end{pmatrix}, \tag{19}$$

where t_{jk}, $j, k = 1, \ldots, D$, are sample variances of pairwise log-ratios between x_j and x_k, i.e.

$$t_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \left(z_{jk}^i - \bar{z}_{jk}\right)^2$$

with

$$z_{jk}^i = \ln \frac{x_{ij}}{x_{ik}}, i = 1, \ldots, n,$$

and

$$\bar{z}_{jk} = \frac{1}{n} \sum_{i=1}^{n} z_{jk}^i.$$

The matrix $\mathbf{T}$ is by construction symmetric with diagonal elements of zero. The elements of the variation matrix can be interpreted in terms of the variability of the ratio between the corresponding parts. Namely, for values t_{jk} close to zero the ratios between x_j and x_k in a given sample are almost constant, so that nearly a perfect proportionality of these parts is achieved. Accordingly, the proportionality of components is considered as the measure of strength of a relationship between two parts.

For theoretical purposes it is important to guarantee that the total variability of a compositional data set does not depend on a particular coordinate representation. The measure of total variability, the *total variance*[25] is defined as a (scaled) sum of all elements of the variation matrix,

$$\mathrm{totvar}(\mathbf{X}) = \frac{1}{2D} \sum_{j=1}^{D} \sum_{k=1}^{D} t_{jk}.$$

Indeed, it can be shown in ref.[4] that totvar(**X**) is equal to the sum of the diagonal elements of the covariance matrix of any ilr coordinates or clr coefficients.

While both center and variation matrix offer an interpretation directly in terms of the compositional parts, they have also some limitations. The center should be considered and interpreted within the Aitchison geometry, where the relative scale matters, thus one should be aware that it is driven by components with small concentrations which contribute at most to the total variability of the compositional data set. Also the strength of the relationship of compositional parts in terms of their proportionality as provided by the variation matrix is not overall acceptable. In many applications, a kind of *correlation* measure which enables to think in terms of *positive* and *negative* relationships would be preferred. This can be achieved for instance with symmetric pivot coordinates, see section ***Further coordinates***. They treat two compositional parts symmetrically and are orthonormal, which is necessary for a geometrically sound computation of the correlation coefficient. These requirements to treat two compositional parts symmetrically and yielding orthonormal coordinates are neither fulfilled with pivot coordinates nor with clr coefficients. The construction of $D(D-1)/2$ symmetric pivot coordinate systems yields the *compositional correlation matrix*,[20] with correlation coefficients contained in the usual interval $[-1,1]$.

2.30.3.2 Visualization Tools

Visualization of compositional data is a delicate point because in principle each component is related through log-ratios to all other parts in a composition. From this perspective it is essentially not possible to visualize single parts. For two-, three- and four-part compositional data a relevant approach is to display them in a fixed representation—in a segment, ternary diagram, and tetrahedron, respectively.[6] Specifically, the ternary diagram is an equilateral triangle $X_1X_2X_3$ such that a composition $\mathbf{x} = (x_1, x_2, x_3)'$ is plotted at a distance x_1 from the opposite side of vertex X_1, at a distance x_2 from the opposite side of vertex X_2, and at a distance x_3 from the opposite side of vertex X_3. The sum of the distances remains constant for any choice of the parts of **x**. However, although a visualization in the original units seems to be beneficial, it might also be quite misleading due to the relative scale of compositional data. Accordingly, compositions near the border possess inherently higher variability than those in the center. This might be easily ignored when looking at the original data with "Euclidean" eyes which might lead to a wrong picture about the data structure, while it would be easily visible when considering any ilr coordinates instead. In fact, frequently compositions with small concentrations in some components form an important source of outliers.

It makes much more sense to look for such visualization tools that make use of the elemental information in compositional data—log-ratios—and their aggregates. The first natural option is to consider univariate graphs of single pairwise log-ratios. However, as there are (up to sign) $D(D-1)/2$ such log- ratios, this option seems to be not much useful for data from chemometrics which are typically high-dimensional. Therefore, an alternative that aggregates some (or all) pairwise log-ratios with a component of interest, as provided by clr coefficients/first pivot coordinates, seems to be preferable. Note that the clr coefficients serve as workhorse for computational issues, but the interpretation of the results and, particularly, possible inference is done by having the first pivot coordinates $z_1^{(l)}$, $l = 1, \ldots, D$, in mind instead. The point is that each of such coordinates is assigned to its own orthonormal coordinate system, and this is not the case with clr coefficients which are interrelating.

Accordingly, a concise picture about the behavior of the single compounds (in the relative sense) can be obtained. One just needs to take into account two possible caveats that might affect the relevance of such visualizations. The first one is linked to possible data quality problems resulting from the imprecision of measurement devices. This affects particularly components with low (absolute) values which are of primary importance due to the relative scale of compositions. Data quality problems include also the presence of values below detection limit (rounded zeros) which are discussed in section ***Zeros*** and possible marginal processes which may drive some of the log-ratios. Fortunately, with higher number of components these effects are usually of less importance.[26] The second problem concerns two- or multiclass data and comes from the fact that such data contain usually just a few log-ratios which can be considered as a source of (bio-)markers (in terms of the original components). While such log-ratios are usually aggregated just in few clr coefficients (or pivot coordinates) which account for the dominance of the original components over an averaged behavior of the other parts, they necessarily affect also other clr variables corresponding to components which cannot be by far considered as markers. For example, while $\ln(x_1/x_2)$ might be one of the "marker log-ratios" leading to a separation of groups of observations (e.g. patients vs. controls) by the first clr coefficient y_1 (and thus indicating x_1 to be one of the markers), the same log-ratio is contained also in y_2 which might then indicate a marker variable as well and could lead finally to an abundance of false positives. Again, this effect can be particularly severe with a small number of components.[27] This needs to be taken into account, and in general it cannot be completely removed without a proper selection of components and/or any kind of their weighting.[28,29] The same also holds for a truly bivariate plotting that is possible in the sense of symmetric pivot coordinates.[20]

Example 1. Commercially available coffee of different origins with different aroma have been analyzed for the concentration of different chemical compounds. The data are accessible in the `R` package `robCompositions` as data set coffee. **Fig. 1** (*left*) shows three of those compounds in a ternary diagram. In this representation, the concentrations are re-scaled to sum up to one. The `coffee` type arabica shows the highest concentration on the compound *furfural*, while robusta shows the lowest concentration. **Fig. 1** (*right*) presents this information in pivot coordinates using Eq. (15). The first coordinate represents all relative information of methylpyrazine to the remaining parts, and the second coordinate relates furfural to dimethylpyrazine. Two arabica coffees are

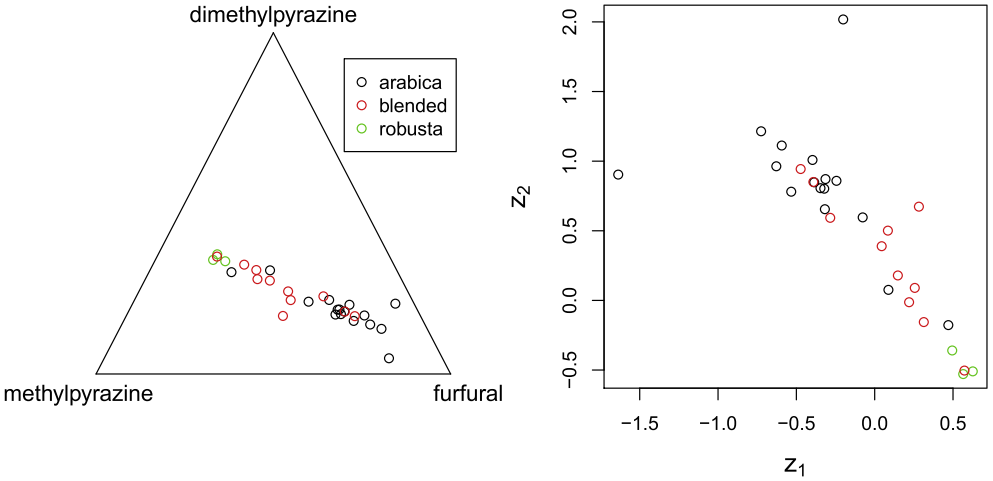

Fig. 1 Ternary diagram (*left*) and coordinate representation (*right*) of three compositional parts of the coffee data.

immediately visible as deviating observations, potential outliers: one has a particularly low concentration on dimethylpyrazine, the other a low concentration on methylpyrazine.

2.30.3.3 Principal Component Analysis

Principal component analysis (PCA) is a key method for dimension reduction.[30] Although it was advocated in the previous sections that working in orthonormal coordinates is essential, PCA with compositional data is usually performed in clr coefficients. The reason is the symmetric form of the coefficients with respect to the original components that can be used to derive an enhanced interpretation of the respective biplot display in terms of pairwise log-ratios. Still, as shown in Kynčlová et al.,[31] PCA in clr coefficients can be closely linked to pivot coordinates in order to follow a consistent approach to a coordinate representation of compositions. Therefore, unlike the lower-dimensional case,[32] for high-dimensional data the following considerations are done just in clr coefficients, where also a robust version of the estimates can be developed.

Denote $\mathbf{y}_i$ as the i-th observation of the clr coefficients, for $i = 1,\ldots, n$, see Eq. (13). Let $\mathbf{C}$ denote the covariance matrix estimated from the matrix of clr coefficients. The spectral decomposition $\mathbf{C} = \mathbf{GLG}'$ is using the diagonal matrix $\mathbf{L}$ containing the eigenvalues in its diagonal, and the matrix $\mathbf{G}$ with the eigenvectors of $\mathbf{C}$. Denoting $\mathbf{t}$ as the estimated center of the clr coefficients, PCA results in a linear transformation,

$$\mathbf{y}_i{}^* = \mathbf{G}'(\mathbf{y}_i - \mathbf{t}) \quad \text{for } i = 1, \ldots n, \tag{20}$$

of the mean-centered clr coefficients into new variables (principal components) such that the first principal component has the largest possible variance (accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The first D–1 variables of the vector of scores $\mathbf{y}_i{}^*$ represent orthonormal coordinates. The last one has zero variance as a direct consequence of the dimensionality of the compositions. The loadings (columns of the matrix $\mathbf{G}$) of the principal components stand for contributions of clr coefficients to the principal components.

In case of high-dimensional data it is more suitable to estimate principal components by singular value decomposition (SVD). SVD decomposes the mean-centered $n \times D$ matrix $\mathbf{Y}$ with clr coefficients $\mathbf{y}_i$ in its rows into three parts,

$$\mathbf{Y} = \mathbf{UDW}', \tag{21}$$

where $\mathbf{U}$ is an $n \times p$ orthogonal matrix containing the left singular vectors, $\mathbf{D}$ is a diagonal matrix of order p containing the (positive) singular values $d_1,\ldots, d_p$, and $\mathbf{W}$ is a $D \times p$ orthogonal matrix containing the right singular vectors. Here, p is the rank of $\mathbf{Y}$, usually $p = \min(n, D–1)$, and it indicates the maximum number of principal components to be considered. So, $\mathbf{p}$ is the minimum of the number of rows and columns of the clr coefficient data matrix $\mathbf{Y}$, the latter reduced by one so that it corresponds to the dimensionality of the compositions. Assume that the left and right singular vectors are sorted according to a decreasing order of the singular values, i.e. $d_1 \geq d_2 \geq \ldots \geq d_p > 0$. When Eq. (21) is rearranged into

$$\mathbf{Y} = (\mathbf{UD})\mathbf{W}' = \mathbf{Y}^*\mathbf{W}', \tag{22}$$

the resulting PCA transformation is obtained. The scores are contained in the matrix $\mathbf{Y}^* = (y_{ij}{}^*)$. The variances of the columns in $\mathbf{Y}^*$ correspond to those of each particular principal component. These variances λ_i, for $i = 1,\ldots, p$, are proportional to the squares of the diagonal elements in the matrix $\mathbf{D}$,

$$\lambda_i = d_i^2/(n-1) \tag{23}$$

Both loadings and scores are used to construct the biplot of compositional data,[33] also called "compositional biplot." Although the purpose of the compositional biplot is the same as for the standard one,[34] i.e. to provide a planar graph that represents a rank-two approximation of both the observations (PCA scores, plotted as points) and variables (loadings, rays) of multivariate data, its interpretation is different: the main interest is in the links (distances between vertices of the rays). Specifically, for the rays i and j (i, $j = 1,\ldots, D$), the link approximates the log-ratio variance $\mathrm{var}\left(\ln\frac{x_i}{x_j}\right)$, which is an element of the variation matrix (19). Hence, when the vertices coincide, or nearly so, then the ratio between x_i and x_j is constant, or nearly so. In addition, directions of the rays indicate where observations with dominance of the corresponding compositional parts are located.

Example 2. The coffee data set from Example 1 is again considered, but the part methylpyrazine is exchanged with the part acetic acid, see ternary diagram in **Fig. 2** (*left*). It can be seen that the robusta coffee samples have very low concentration on acetic acid, and thus they clearly form outliers. This is also visible in the coordinate presentation in **Fig. 2** (*right*). The two lines in the plots show the directions of the first principal component (PC1), once estimated based on the classical sample covariance matrix, as described in the text above, and once robustly based on a robust estimate of location and covariance.[35] The outlier group strongly attracts the direction of PC1, while for the robust solution, these atypical observations are downweighted.

When using all available six compositional parts from the coffee data set, PCA can be used for dimension reduction. The biplots of the first two PCs are shown in **Fig. 3**: the left plot for classical PCA, the right plot for robust PCA. The types of coffees (arabica, blended, robusta) are abbreviated by their first letter, and the corresponding PCA scores are shown in the biplots, together with the rays for the parts. Again it is visible that PC1 for classical PCA has been attracted by the robusta group. The robust biplot allows to get a better insight into the multivatiate data structure of the data majority.

2.30.3.4 Outlier Detection

Outlier detection is often considered as an exploratory method to learn something about the data structure. Observations deviating from the data majority can be identified, and the reason for their outlyingness can be investigated. Sometimes there are measurement or reporting errors, or the observations are inconsistent for any other reason, which is usually an important message. Since compositional data are multivariate data, outlier detection needs to be carried out in a multivariate sense.

Most statistical outlier detection methods rely on the usual Euclidean geometry, because they are based on distances and/or covariance estimation. Thus, as a first step, the $n \times D$ compositional data matrix $\mathbf{X}$ needs to be expressed in coordinates, e.g. using Eq. (15). This yields the matrix $\mathbf{Z}$ with n rows and D–1 columns. Denote the observations (rows) as $\mathbf{z}_i$, for $i = 1,\ldots, n$. A common tool for multivariate outlier detection is to use Mahalanobis distances of each observation $\mathbf{z}_i$ to the center $\mathbf{t}$ with respect to the covariance matrix $\mathbf{C}$, i.e.

$$\mathrm{MD}(\mathbf{z}_i) = \left((\mathbf{z}_i - \mathbf{t})'\mathbf{C}^{-1}(\mathbf{z}_i - \mathbf{t})\right)^{1/2} \quad \text{for } i = 1,\ldots,n. \tag{24}$$

It is essential that $\mathbf{t}$ and $\mathbf{C}$ are robust estimates, and there are several proposals available in the literature, such as the Minimum Covariance Determinant (MCD) estimator.[35] A common cutoff value for identifying outliers is the value $\sqrt{\chi^2_{D-1;0.975}}$; observations exceeding this threshold are considered as multivariate outliers.

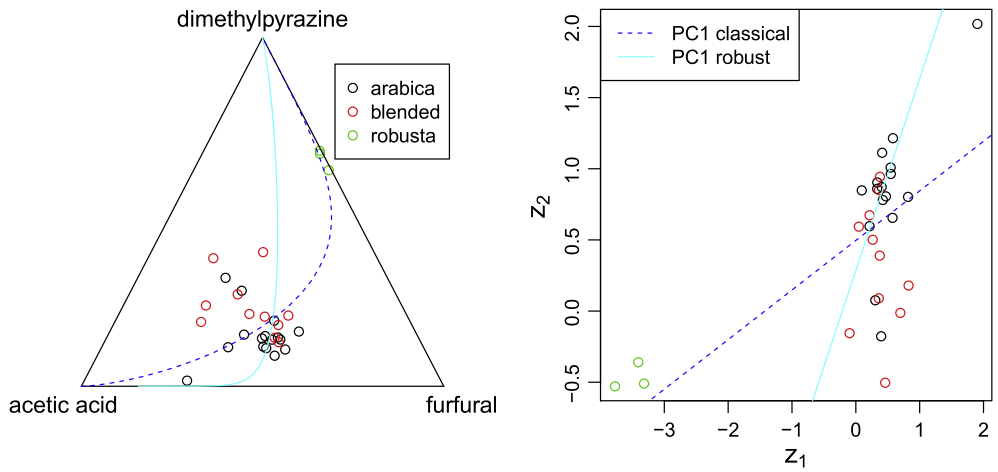

Fig. 2 Ternary diagram (*left*) and coordinate representation (*right*) of three compositional parts of the coffee data. The lines indicate the directions of the first classical and robust principal component.

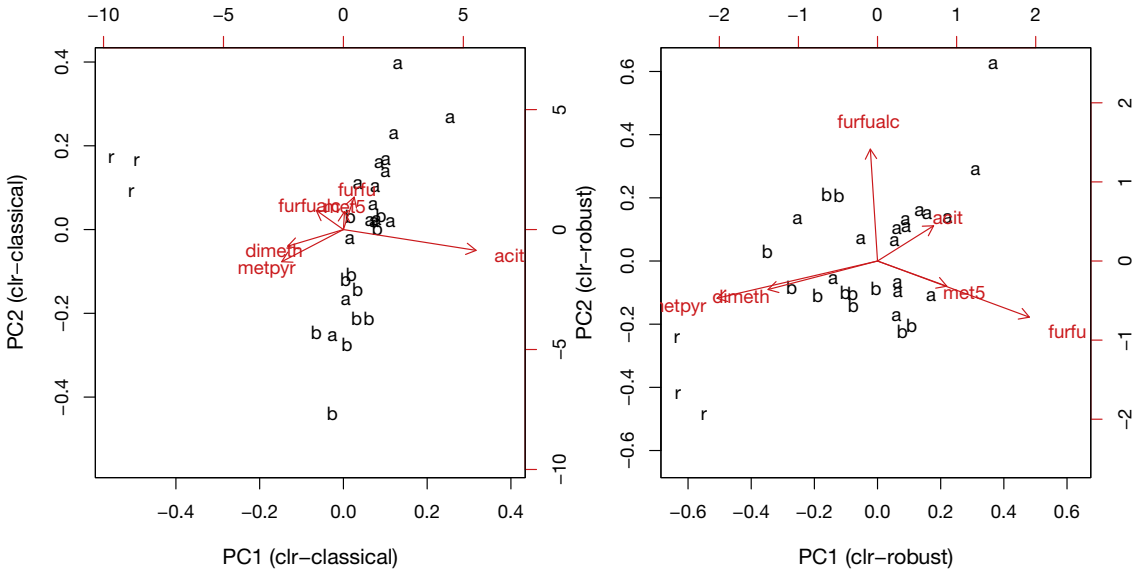

Fig. 3 Classical (*left*) and robust (*right*) PCA of all six compositional parts of the coffee data. Shown are the biplots, with rays representing the compositional parts, and letters for the coffee type (arabica, blended, robusta) representing the scores.

The MCD estimator is affine equivariant, which is convenient since any ilr coordinates could be used without changing the resulting Mahalanobis distances.[36] Note that this estimator would not work for clr coefficients, because a full rank input matrix is required.

For high-dimensional data, in particular when the number of observations is smaller than the number of compositional parts, the above approach for outlier detection would not work because of singularity. In this case there are other proposals available, e.g. Filzmoser et al.[37] These methods are typically no longer affine equivariant, and thus the choice of the ilr coordinates would matter. It is thus recommended to use these tools rather for data exploration than in terms of a strict statistical test for outliers. On the other hand, these tools are still very valuable to identify unusual observations, which very likely would also be detected with a different choice of the ilr coordinates.

Example 3. The same compositional parts from the coffee data set are used as in Example 1. The ternary diagram and a coordinate presentation is shown in Fig. 4. These plots also include the so-called 97.5% tolerance ellipse. This ellipse is formed by all $\mathbf{z}$ for which $MD(\mathbf{z}) = \sqrt{\chi^2_{2;0.975}}$ is fulfilled. Here, the location and covariance for the Mahalanobis distance are estimated robustly by the MCD estimator, see Eq. (24). In case of bivariate normal distribution, this ellipse would contain the innermost 97.5% of the data. The ellipse as constructed here corresponds to the region separating regular observations from outliers. Only in the coordinate representation, the ellipse appears correctly as an ellipse (usual Euclidean geometry), while the visual impression in the ternary diagram (simplex) is different.

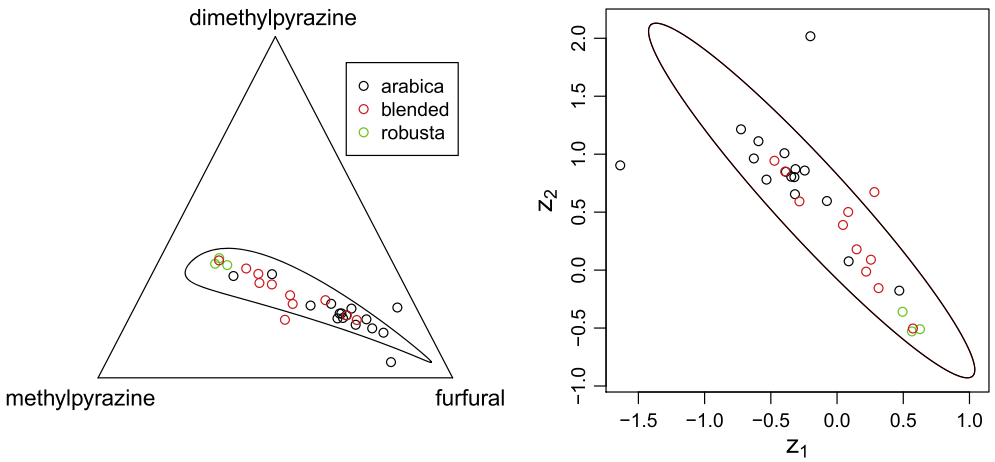

Fig. 4 Ternary diagram (*left*) and coordinate representation (*right*) of three compositional parts of the coffee data. Observations inside the ellipses are regular observations, outliers are outside.

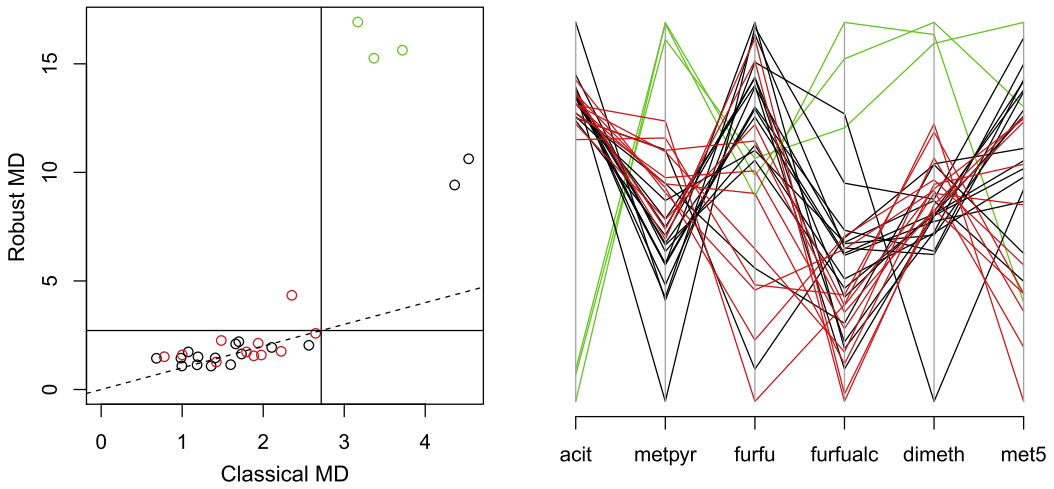

Fig. 5 Distance-distance plot for the coffee data for outlier diagnostics (*left*), and parallel coordinate plot for the interpretation of the outlyingness (*right*).

For the following analysis, all six compounds of the coffee data set are used for multivariate outlier detection. Mahalanobis distances are computed based on classical and robust (MCD) estimates, and presented in Fig. 5 (*left*). The horizontal and vertical line in this so-called distance-distance plot represent the outlier cutoff values. Observations exceeding these values would be revealed as multivariate outliers. However, only for the robust distance this diagnostics is reliable. The samples from the robusta group are clearly identified as outliers, even with the classical distance. Also two arabica samples are atypical, and one blended sample is only identified as outlier with the robust method. Fig. 5 (*right*) shows a parallel coordinate plot. The coordinates are clr coefficients of the compounds, and each line represents one coffee sample. One can see that indeed the data structure of the robusta samples are different, and in this plot it also becomes clear in which compounds they differ.

2.30.4 Linear Regression

Linear regression represents a popular tool to model the (linear) relationship between one or several response variables and one or more explanatory variables. In chemometrics, concentrations or spectra usually play the role of the explanatory variables, while the response is a given (real) output variable. From the compositional perspective this requires a proper coordinate representation of the covariates that enables for a reasonable estimation of the regression parameters and a corresponding statistical inference. Here the clr coefficients can be used again as workhorse for computational issues, see, e.g., Bruno et al.,[38] but the regression model itself is formulated rather in terms of ilr coordinates, preferably the pivot ones. While in lower dimensions the regression parameters can be estimated using the traditional least-squares method (or a robust counterpart), for high-dimensional data an appropriate generalization is required. In this section the focus is on the case with a non-compositional response and compositional explanatory variables. This can be easily extended to the multivariate case, i.e. several non-compositional response variables. Other cases where the responses are compositional, with non-compositional or also compositional explanatory variables, are treated in Filzmoser et al.[6]

2.30.4.1 Methods in Lower Dimension

When D-part compositions $\mathbf{x}_i = (x_{i1},\ldots,x_{iD})'$ for $i = 1,\ldots, n$ are expressed in ilr coordinates, or more specifically, in any of the D pivot coordinate systems $z_1^{(l)}, \ldots, z_{D-1}^{(l)}$, it is possible to form a standard multiple regression model that can be used for further estimations, e.g. by the least-squares (LS) method. When n measurements of the covariates are taken together with those of the real response variable Y with observations $\mathbf{y} = (y_1,\ldots, y_n)'$, the resulting models can be written as follows,

$$y_i = b_0^{(l)} + z_{i1}^{(l)} b_1^{(l)} + \cdots + z_{i,D-1}^{(l)} b_{D-1}^{(l)} + e_i, i = 1,\ldots,n; l = 1,\ldots,D, \tag{25}$$

or in matrix form as

$$\mathbf{y} = \mathbf{Z}^{(l)}\mathbf{b}^{(l)} + \mathbf{e}, \tag{26}$$

where the first column of the $n \times D$ matrix $\mathbf{Z}^{(l)}$ is filled with ones, the regression coefficients $\mathbf{b}^{(l)} = (b_0{}^{(l)}, b_1{}^{(l)},\ldots, b_{D-1}^{(l)})'$, and the error terms $\mathbf{e} = (e_1,\ldots, e_n)'$. It can be shown[15] that by using rotations between the different pivot coordinate systems, the LS estimates of the parameters $b_0{}^{(l)} \equiv b_0$ are the same for all $l = 1,\ldots, D$. The same holds for the prediction of the response variable and for further model characteristics. The list includes also the residual sum of squares, given as sum of squared differences between observed and predicted values of the response, the F-statistic and the coefficient of determination.[39] In addition to the absolute term parameter,

also the parameters $b_1^{(l)}$ are of particular interest due to the interpretation of the first pivot coordinates. Because the coordinates $z_2^{(l)}$, …, $z_{D-1}^{(l)}$ fully represent the remaining parts $x_2^{(l)}$, …, $x_D^{(l)}$, they cannot be avoided from the model, although the corresponding regression parameters are rather rarely taken for interpretation purposes.

The estimates of the parameters b_0, $b_1^{(l)}$, …, $b_1^{(D)}$ together with their further characteristics (standard errors, values of the *t*-statistics and the respective *p*-values) are usually jointly presented in one table, as if they all would result from one regression model. However, it is important to realize that the outputs come from D regression models, and thus attempts like considering the *F*-statistic for joint significance testing of the parameters $b_1^{(l)}$, …, $b_1^{(D)}$ should be avoided.

Example 4. The `R` package `chemometrics` contains the data set `ash`, which consist of 99 ash samples originating from different biomass. The response variable is the softening temperature (SOT) which should be predicted using the information of eight oxides, see also Varmuza and Filzmoser.[40] Since nine observations contain zeros, they are omitted here for simplicity. For the remaining data, a robust regression model is estimated using LTS-regression.[41] As described above, the statistical inference is obtained by estimating a model for each of the D pivot coordinate systems, and using the inference information from only the first coordinate in each system. The estimated regression coefficients and the statistical inference information is presented in **Table 1**. Accordingly, one can see that the coordinates for P_2O_5, CaO and K_2O are significant for predicting SOT, the latter two with negative sign. The R^2 measure (coefficient of determination) indicates a reasonable model fit.

Fig. 6 shows the regression diagnostic plot, with robust Mahalanobis distances on the horizontal axis, and robust standardized residuals on the vertical axis, together with cutoff values as lines, see also Rousseeuw and Van Driessen.[41] The lower and upper right regions contain observations which are leverage points, i.e. in a classical least-squares regression they could have a strong leverage effect. Thus, using robust regression protected against these outliers, which are downweighted for the estimation of the regression parameters.

Fig. 7 contrasts the measured with the predicted response variable SOT. Red points are observations which are downweighted in the robust regression model (outliers). One can see that these are mainly values with high SOT, where the multivariate information of the oxides must be somewhat different.

2.30.4.2 Methods in Higher Dimension

In higher dimensions the same regression model (26) can be considered, but for the estimation of the regression parameters the LS method cannot be used any more, because it requires full column rank of the matrix $\mathbf{Z}^{(l)}$, which is not the case if the number of explanatory variables exceeds the number of observations. A way out is to employ partial least squares (PLS) regression instead which aims at finding a linear regression model by projecting the explanatory variables to a new space, obtained by maximizing the covariance between the covariates and the response.[42] Formally, consider a weight vector w, and the projection $\mathbf{t} = \mathbf{Zw}$. Without loss of generality, the matrix $\mathbf{Z} = \mathbf{Z}^{(1)}$ is considered here. Then the maximization problem can be written as

$$\max_{\|\mathbf{t}\|=1} \operatorname{cov}(\mathbf{t}, y) = \max_{\|\mathbf{Zw}\|=1} \operatorname{cov}(\mathbf{Zw}, y). \tag{27}$$

The solution of the maximization problem is formed by the first weight vector $\mathbf{w}_1$, and the resulting first score vector $\mathbf{t}_1 = \mathbf{Zw}_1$. For the subsequent score vectors, orthogonality constraints to the previous score vectors are imposed, i.e., $\mathbf{t}_j'\mathbf{t}_k = 0$ for $1 \le j \le k \le a$, where a is a given number of components to be derived. Finally, the weight matrix $\mathbf{W}$ and the score matrix $\mathbf{T}$ are derived, with the weight vectors $\boldsymbol{w}_j$ and score vectors $\mathbf{t}_j$, respectively, in the columns, for $j = 1,\ldots, a$. Instead of using the original covariates in a regression problem $\mathbf{y} = \mathbf{Zb} + \boldsymbol{\varepsilon}$, the scores are used in the model $\mathbf{y} = \mathbf{T}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, i.e. the regression is done on "partial" information of the covariates. In the latter model, the classical LS estimator can be employed because a is usually much smaller than D, yielding $\widehat{\boldsymbol{\beta}}$. Since $\mathbf{T} = \mathbf{XW}$, a back-transformation gives the estimation of the regression coefficients in the original space as $\widehat{\mathbf{b}} = \mathbf{W}/\widehat{\boldsymbol{\beta}}$. Similarly as for PCA, also here it is assumed that both covariates, collected in the matrix $\mathbf{Z}$, and the response $\mathbf{y}$ are mean-centered. Accordingly, the absolute term is omitted from the regression model (26) and the matrix $\mathbf{Z}$ is just of dimension $n \times (D-1)$.

Table 1 Regression coefficients and statistical inference from robust regression with the ash data set.

	Estimate	*Std. error*	t-*Value*	*Pr(>|t|)*	
Intercept	1343.44	54.51	24.645	< 2e-16	***
ilr(P_2O_5)	101.15	19.08	5.302	1.27e-06	***
ilr(SiO_2)	21.36	11.25	1.898	0.06148	.
ilr(Fe_2O_3)	40.34	21.79	1.851	0.06840	.
ilr(Al_2O_3)	6.27	17.86	0.351	0.72658	
ilr(CaO)	−73.69	22.68	−3.249	0.00178	**
ilr(MgO)	31.66	28.68	1.104	0.27338	
ilr(Na_2O)	−14.27	11.78	−1.212	0.22963	
ilr(K_2O)	−139.55	19.77	−7.060	9.70e-10	***

Significance codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1.

Residual standard error: 89.03 on 66 degrees of freedom Multiple *R*-squared: 0.7611, Adjusted *R*-squared: 0.7372; *F*-statistic: 31.86 on 7 and 70 DF, *P*-value: < 2.2e-16.

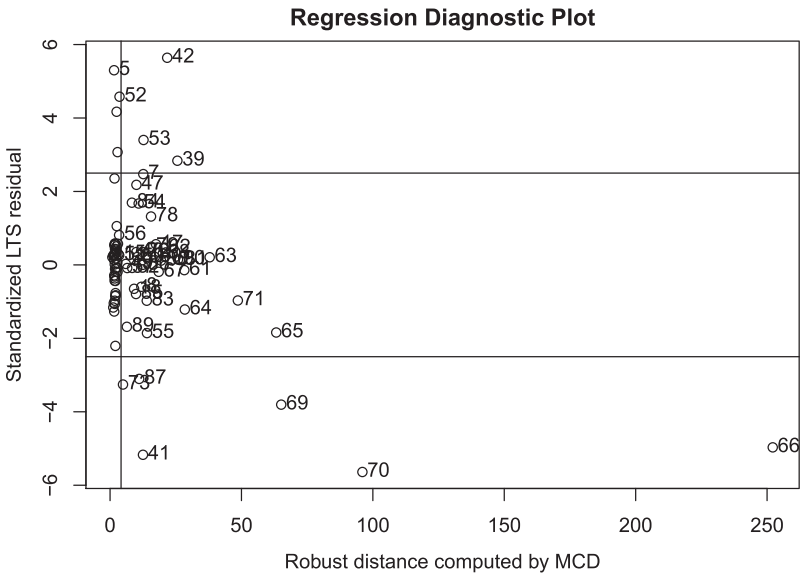

Fig. 6 Regression diagnostic plot for the resulting robust regression model with the ash data set.

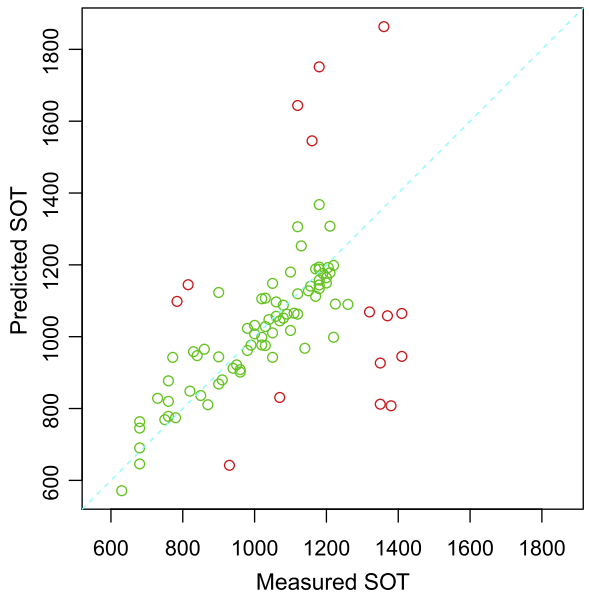

Fig. 7 Measured versus predicted softening temperature (SOT) using the robust regression model with the ash data set; outliers are indicated by red color.

There are several algorithms for solving the PLS problem, such as Kernel PLS, NIPALS, SIMPLS, or O-PLS.[40] Since each additional score vector covers new variability, it is preferable for prediction purposes to have uncorrelated scores like for PCA. However, in case of PLS regression it is redeemed by the fact that the resulting scores represent coordinates which are not orthogonal to each other.

As mentioned above, PLS regression contains inherently a projection to a new space where the regression modeling is performed. Therefore, PLS is also used for dimension reduction, where the resulting loadings and scores are displayed in form of a biplot, similarly as for the case of PCA. For this purpose, the PLS computations are done in clr coefficients and their link to the first pivot coordinates (16) is utilized for interpretation purposes.

2.30.5 Linear Classification

For classification it is assumed that next to the multivariate data information also a group label for the membership of each observation to a particular data group is available. Generally, there might be $G \geq 2$ different groups available, and the groups may refer to

different classes of patients, to the region of origin of the samples, etc. The task is to use training data for the estimation of a function that best possibly separates the multivariate training data into the distinct groups, e.g. by minimizing an overall misclassification error. Afterwards, this classifier can be used to classify new test data. In this section, linear classification functions are established, which means that linear combinations are considered, with specific coefficients that need to be estimated based on the training data. As already argued previously, the compositional input data cannot be directly used for this purpose, but they need to be mapped into the standard Euclidean space for which the traditional linear classifiers are designed.

2.30.5.1 Methods in Lower Dimension

The focus here is on the two-group case, i.e. $G = 2$. Extensions to the multigroup case are possible and follow the same principle. The first step is to express the compositional data matrix in coordinates, say as matrix $\mathbf{Z}$ of dimension $n \times (D-1)$. This can be done by making use of Eq. (15), but also any other choice of orthonormal coordinates is possible. In fact, Filzmoser et al.[43] have shown that for the discriminant analysis methods described in the following, any choice of orthonormal coordinates would lead to the same decision rule. Note, however, that clr coefficients would not be appropriate: the rules mentioned below require an estimation of the inverse covariance matrix, and clr coefficients end up in a singular covariance matrix for which the inverse does not exist.

2.30.5.1.1 Fisher discriminant rule

Since the group membership is known for the training data $\mathbf{Z}$, it is possible to split up this matrix into a matrix $\mathbf{Z}_1$ consisting of the n_1 observations of the first group, and a matrix $\mathbf{Z}_2$ with the remaining n_2 observations ($n_1 + n_2 = n$). Denote the sample mean vectors of these matrices as $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$, respectively. Consider a vector $\mathbf{a} = (a_1,\ldots, a_{D-1})'$, and the linear combinations $\mathbf{y}_1 = (y_{11},\ldots,y_{1n_1})' = \mathbf{Z}_1'\mathbf{a}$ and $\mathbf{y}_2 = (y_{21},\ldots,y_{2n_2})' = \mathbf{Z}_2'\mathbf{a}$. Further, let $\bar{y}_1 = \bar{\mathbf{z}}_1'\mathbf{a}$ and $\bar{y}_2 = \bar{\mathbf{z}}_2'\mathbf{a}$. The Fisher discriminant rule aims at identifying a vector a such that the expression

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y} \tag{28}$$

is maximized, where

$$s_y^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}{n_1 + n_2 - 2} \tag{29}$$

is the estimated pooled variance of both groups. This means that the multivariate data are projected on one dimension, such that the projected group means are as much separated as possible with respect to the variance of the projected values.

Eq. (29) can also be written as

$$s_y^2 = \mathbf{a}'\mathbf{S}_{pooled}\mathbf{a},$$

with the estimated pooled covariance matrix

$$\mathbf{S}_{pooled} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

and the sample covariances $\mathbf{S}_1$ of $\mathbf{Z}_1$ and $\mathbf{S}_2$ of $\mathbf{Z}_2$, respectively.

One can show[39] that criterion (28) is maximized for

$$\mathbf{a} = \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2) \tag{30}$$

The decision rule is then as follows:

- Consider a new test set observation $\mathbf{x} = (x_1,\ldots x_D)'$, which is a composition with the same compositional parts as the training data.
- Use the same representation in coordinates that has been used for the training data, yielding $\mathbf{z} = (z_1,\ldots, z_{D-1})'$.
- Compute $\mathbf{a}$ according to Eq. (30) from the training data, the projected group means $\bar{y}_1$ and $\bar{y}_2$, and the projected value $y = \mathbf{z}'\mathbf{a}$ for the test set observation.
- If $y - \bar{y}_1$ is smaller than $y - \bar{y}_2$, assign the test set observation to group 1, otherwise to group 2.

2.30.5.1.2 Bayesian discriminant rule

As mentioned in section *Fisher discriminant rule*, the compositions first need to be expressed in orthonormal coordinates, and the specific choice would not alter the resulting decision rule. Further, it is assumed that prior probabilities p_1 and p_2 ($p_1 + p_2 = 1$) can be assigned to the two groups, indicating the probability that an observation originates from the specific group. It is also assumed that the groups are multivariate normally distributed, with specific means and covariances. In case of linear discriminant analysis, a further assumption is that the covariances of both groups are identical, and thus the groups only differ in their mean. Making use of the Bayesian theorem allows to express the posterior probabilities in terms of the prior probabilities and the group density functions. The decision boundary is the value where the posterior probabilities of an observation are equal for both groups, which

means that their log-ratio is zero. It turns out[39] that this log-ratio can be expressed as $\delta_1(\mathbf{z})$–$\delta_2(\mathbf{z})$, for an observation $\mathbf{z}$ (expressed in coordinates), and with the linear discriminant scores

$$\delta_j(\mathbf{z}) = \left(\mathbf{z} - \frac{1}{2}\bar{\mathbf{z}}_j\right)' \mathbf{a} - \ln p_j \quad \text{for } j = 1, 2. \tag{31}$$

Here,

$$\mathbf{a} = \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_2), \tag{32}$$

which is the same solution as for the Fisher method, see Eq. (30) if again the classical estimators for the parameters are used.

The decision rule is thus as follows:

- Consider a new test set observation $\mathbf{x} = (x_1,\ldots x_D)'$, which is a composition with the same compositional parts as the training data.
- Use the same representation in coordinates that has been used for the training data, yielding $\mathbf{z} = (z_1,\ldots, z_{D-1})'$.
- Compute $\mathbf{a}$ according to Eq. (32) from the training data, and the group means $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$.
- Estimate the prior probabilities p_1 and p_2, e.g. by the proportions of the observations of the training data in the groups, and compute the linear discriminant scores according to Eq. (31).
- If $\delta_1(\mathbf{z})$ is bigger than $\delta_2(\mathbf{z})$, assign the test set observation to group 1, otherwise to group 2.

2.30.5.1.3 More than two groups

Both the Fisher approach as well as the Bayesian discriminant rule can be extended to the case of more than two groups, see, e.g. Filzmoser et al.[6] For the Fisher discriminant analysis method, the *within groups* covariance matrix needs to be estimated, e.g. by a pooled version of all group covariances. Moreover, a matrix describing the variation *between the groups* is needed, reflecting the squared distances of the group centers to the overall center. Maximizing the *between* relative to the *within* covariance leads to a projection space which is most informative for the group separation and is thus typically used to visualize the classification problem.

Example 5. The well known olive oil data set with measurements of fatty acids in olive oils originating from different regions in Italy is available as data set `olives` in the `R` package `classifly`. In order to avoid problems with zeros, those observations are simply deleted from the data. Also, one variable contains measurement problems with values below a detection limit, and thus this variable (eicosenoic) is omitted as well. The data set used here has 535 observations, originating from southern Italy, Sardinia, and northern Italy, and seven fatty acids. The data are treated as compositional data because the interest is in the ratios between the fatty acids rather than directly in the concentration data. Thus, pivot coordinates are computed and used for Fisher discriminant analysis with three groups. The discriminant functions will be estimated based on a randomly selected training set of about 2/3 of the observations, and evaluated on the remaining test data.

Fig. 8 presents the plot of the discriminant scores for the training data. Note that a three-group problem reduces to only two discriminant scores. The colors of the symbols refers to the true group membership, while the symbols refers to the predicted group

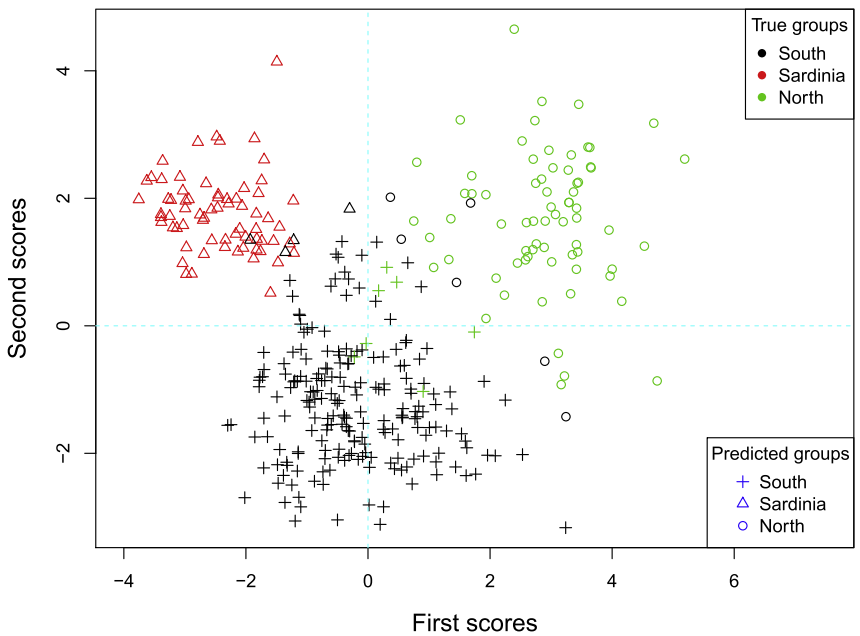

Fig. 8 Projection of the olive oil training data into the space derived by Fisher discriminant analysis.

Table 2 Classification result for the training data of the olive data set using Fisher discriminant analysis.

True group	*Predicted group*		
	South	*Sardinia*	*North*
South	202	4	6
Sardinia	0	68	0
North	7	0	70

membership. There are only few misclassified observations, and the details are listed in **Table 2**. This results in a misclassification rate of 4.8% for the training data.

The test data can be projected into the space derived from Fisher discriminant analysis, resulting in a visualization of the test set samples (**Fig. 9**) and in a classification table (**Table 3**). Especially samples from northern Italy seem to be inconsistent and are thus misclassified, but the misclassification rate is still only 9.0%.

2.30.5.2 Methods in Higher Dimension

Neither the Fisher discriminant rule nor the Bayesian discriminant rule can be used for classification with high-dimensional data, where the number of observations is lower than the number of variables, because they both require the inverse of the pooled covariance matrix which cannot be achieved in such a case. One possible approach is to apply a procedure which contains inherently dimension reduction instead. This is provided by the partial least squares regression discriminant analysis (PLS-DA) method. This method can be considered as a special case of PLS regression described in section ***Methods in higher dimension***.[40] Here we focus on the two-group case which frequently occurs in practice. Accordingly, the vector $\mathbf{y}$ in Eq. (26) consists of zeros and ones which correspond to samples from either categories. The algorithm is optimized for the balanced case. This means that there is (raughly) the same amount of members in both categories. After centering of the covariates and the response which is common with PLS regression (see section ***Methods in higher dimension***), the categories are represented by negative and positive constants instead of zeros and ones. This is advantageous for the interpretation of regression coefficients; accordingly, the marker variable can be assigned to either of the groups by considering the sign of the respective coefficient.

Similarly as in case of LS regression from section ***Methods in lower dimension***, the output of the regression procedure are estimates of the parameters $b_1{}^{(1)}, \ldots, b_1{}^{(D)}$ denoted as $\widehat{b}_1^{(1)}, \ldots, \widehat{b}_1^{(D)}$ (remind that we are working with centered data, therefore the absolute term parameter is omitted). However, the statistical inference cannot be done without having the exact distribution of test statistics, and thus an approximative approach is needed. Here the bootstrap procedure is described which works well for both the balanced and unbalanced cases.[44] The idea is to draw random samples with replacement from each group of the original data, where the bootstrap group samples have the same size as the original groups. This results in a bootstrap data set for the explanatory variables and the response, where PLS-DA is applied to estimate the parameters. Repeating this procedure many times allows to estimate the

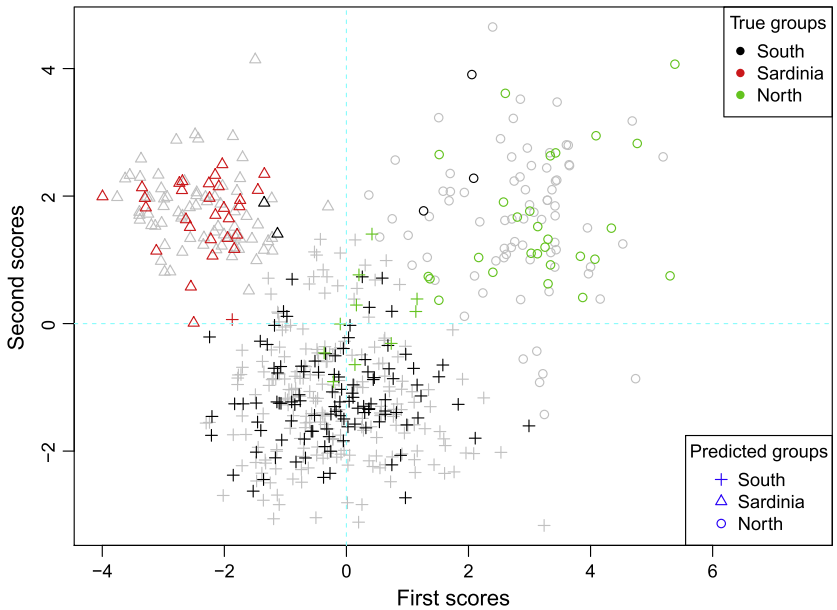

Fig. 9 Projection of the olive oil test set data into the space derived by Fisher discriminant analysis. The scores for the training data are shown in *gray*.

Table 3 Classification result for the test data of the olive data set using Fisher discriminant analysis.

True group	*Predicted group*		
	South	*Sardinia*	*North*
South	106	2	3
Sardinia	1	29	0
North	10	0	27

variability of the regression parameters. The standardized regression estimates are then obtained by dividing the regression parameters of the original data by the respective estimated standard deviations $s_1,\ldots, s_D$ from the bootstrap, and they can be compared with quantiles of the standard normal distribution.

Accordingly, if the standardized estimates $\widehat{b}_1^{(1)}/s_1,\ldots,\widehat{b}_1^{(D)}/s_D$ are lower than the $\alpha/2$ or higher than the $(1-\alpha/2)$ quantile of the standard normal distribution (usually $\alpha = 0.05$ is taken), the corresponding parameter is considered to be significantly different from zero, and thus the respective variable contributes to the discrimination task.

In order to reduce the risk of false positives, a Bonferroni correction is usually applied, resulting in an adjusted a-level of significance, $\alpha_{\text{adj}} = \alpha/D$. Since each of the covariates comes from another coordinate system, this correction is not necessarily needed from a theoretical perspective, despite it is recommended for the above mentioned reason.

2.30.6 Data Preprocessing

In chemometrics, the observations typically have one important step in common. The raw data matrix, which may consist of, for example, chromatographic peak heights, or baseline corrected spectroscopic intensities, are scaled or reduced in size prior to statistical processing. The reason is usually that occasional large values, e.g. peaks in GCMS or NMR spectroscopy that are very high relative to others in a few samples would otherwise dominate the analysis if not properly transformed or scaled. Frequently also the argument occurs to symmetrize the distribution of concentrations which is commonly right skewed. Another problem typical for areas dealing with biological material like metabolomics is the so called *size effect* (see, e.g., ref.[27]), associated with a different sample volume and/or sample concentration. Due to the size effect, the true signal is unobservable, but what is observed is a signal that is multiplied by a constant, and the constants in general differ for different signals.

2.30.6.1 Normalization and Scaling

There are numerous approaches how to normalize data in chemometrics, see Brereton,[45] or Walach et al.[46] for an exhaustive overview in omics sciences. In the sequel those methods are listed which are closely connected to the inherent properties of compositional data.

The common aim of normalization techniques is to *transform* the original compounds in whatever units to dimensionless variables that reduce undesired effects as described above. Two prominent examples of this concept are AUC normalization and the PQN transformation. The AUC normalization aims to normalize a group of signals with peaks by standardizing the area under the curve (AUC) within a sample to the median, mean or any other proper representation of the amount of dilution. For the case of the mean, one gets for the i-th observation $\mathbf{x}_i = (x_{i1},\ldots, x_{iD})'$ with D components a new signal (observation)

$$\mathbf{x}_i^{AUC} = \left(\frac{x_{i1}}{\frac{1}{n}\sum_{j=1}^{D} x_{ij}},\ldots,\frac{x_{iD}}{\frac{1}{n}\sum_{j=1}^{D} x_{ij}}\right)' \tag{33}$$

By doing so, the resulting data matrix with the signals in the rows would have the same row sums (n), what is somehow undesirable for many multivariate methods. Similarly, when simply the sum of the compounds would be taken instead of mean, row sums 1 would be reached; in that case this is referred to total sum normalization (TSN). Therefore, taking the median for each signal might seem to be a way out, and the row sums in the resulting data set vary again. The median is used in a more sophisticated way also in probabilistic quotient normalization (PQN). Here for the i-th fingerprint, the size effect s_i is estimated based on the median of the ratios of the elements of $\mathbf{x}_i$ and the corresponding elements of a preselected reference signal $\mathbf{x}_{ref} = (x_1^{ref},\ldots,x_D^{ref})'$. Accordingly,

$$\mathbf{x}_i^{PQN} = \left(\frac{x_{i1}}{s_i},\ldots,\frac{x_{iD}}{s_i}\right)', \tag{34}$$

where

$$s_i = \text{median}\left\{\frac{x_{i1}}{x_1^{ref}},\ldots,\frac{x_{iD}}{x_D^{ref}}\right\}$$

The resulting signal x_i^{PQN} still keeps the original units by the hope of getting rid of both size effect and any side-effect-imposed constant sum constraint. Therefore it is used with increasing popularity.[27]

Both AUC normalization and PQN transformation have something in common: they stand for a proper *representation* of information contained in ratios. In other words, by multiplying each sample by a proper positive constant, a PQN transformed composition can be expressed with the unit sum constraint of components, and vice versa. Therefore, despite being more or less sophisticated, neither of these transformations can avoid possible problems when processing statistically the original compositions: their relative scale is ignored and tools like PCA or PLS-DA designed to work with real multivariate data are applied to positive-valued variables enriched with the scale invariance property. Of course, the multivariate information contained in log-ratios can be more or less effectively processed by either proper choice of log-ratio coordinates or an efficient statistical tool for pattern recognition, but this is already another issue.

A further important aspect of data preprocessing is scaling (standardization) which involves first mean centering and then dividing each column (or variable) by its standard deviation. Scaling ensures that the variation in each variable has an influence that is approximately equal on the result of the statistical analysis. In case of compositional data expressed in log-ratio coordinates usually only centering is performed. Applying log-ratios themselves usually does the work for which dividing variables by their respective standard deviations is aimed; but, more importantly, scaling would destroy the possibility of rotating the data in order to get another ilr coordinate representation. This property is important for most methodological developments with compositional data including PCA and PLS modeling.

2.30.6.2 Missing Values

Although missing values are rather a typical phenomenon in survey data, they occur also in chemometrics[45] and refer to a situation where no information was recorded for a specific variable in a specific sample. The reason may have been because the instrument was not working properly on a particular day which sometimes happens, for example, in environmental research. Removing such a sample is not tenable when dealing with high-dimensional data because too much information would be lost by doing so. The problem with missing values becomes even more severe with compositional data because the missings would expand into all log-ratio coordinates where the respective components are contained.

If a sample is retained with a small number of missing values, then one possible approach is to replace each missing value with the geometric mean of the component over the entire dataset (see section ***Statistical data summaries***). However, this approach becomes quickly unreliable with an increasing amount of missing values because such an imputation completely ignores the multivariate data structure and reduces artificially the variation of the components. A possible way out with high-dimensional data is to apply *k*-nearest neighbor imputation where the idea is to use a distance measure (here the Aitchison distance) for finding the *k* most similar observations to a composition containing missings, and to replace the missings by using the available variable information of the neighbors.[47] For imputing a missing part of a composition the median of the corresponding cells of the *k* nearest neighbors can be used. However, the cells first need to be adjusted according to the overall size of the parts, because the Aitchison distance is the same for any representation of the input compositions. Alternatively, it would also be possible to adapt the model-based algorithm introduced in the next section for missing value imputation.

2.30.6.3 Zeros

Zero values of components are not allowed within the log-ratio methodology of compositional data, because either a division by zero or the logarithm of zero are not defined. Another, geometrical, reason is that in the context of the Aitchison geometry zeros play the role of infinity. This is closely related again to the relative scale of compositional data. Accordingly, the closer the concentrations are to zero, the more log-ratios of them with other parts exceed any bound (either in the positive or negative sense). This, however, embarrass practical data analysis where zeros naturally occur, and therefore procedures that are able to deal with zeros are needed.

In chemometrics zeros usually stand for *values below the detection limit* of a measurement device (also identified with rounded zeros) rather than completely absent values which are referred to as *structural zeros* in the compositional context. Rounded zeros can be imputed by relevant values below the detection limit, and then one can proceed further with the analysis. The simplest way to deal with rounded zeros is to replace them with a constant value below the detection limit. In compositional data analysis, the value 2/3 of the detection limit is recommended.[48] However, in that case the same problem occurs like when imputing missing values by the mean of the respective component values. The relations between the variables are completely ignored and the variation of the data set is decreased artificially. Moreover, the effects of wrong imputation are even more severe due to relative scale of the compositions.

Therefore, a model-based approach for imputation of values below the detection limit is highly recommendable. In Templ et al.[49] such an iterative procedure is proposed based on censored regression using PLS regression, where a set of pivot coordinate systems $z_1^{(l)}, \ldots, z_{D-1}^{(l)}$ needs to be employed sequentially for $l \in \{1, \ldots, D\}$ in order to perform the imputation for each of the original compositional parts. As a starting point, zeros are imputed by 2/3 of the respective detection limits. In the next step, the compositional parts are rearranged (without loss of generality) such that x_1 includes the highest amount of missings, x_2 the second highest, and so on. Thus, when performing the PLS regression of z_1 on $z_2, \ldots, z_{D-1}$, only z_1 will be influenced by the initialized zeros in x_1, but not the remaining pivot coordinates. The idea of the procedure is thus to iteratively improve the estimation of the rounded zero values. After the regression of z_1 on $z_2, \ldots, z_{D-1}$ the results are back-transformed to the space of the original data, the extended

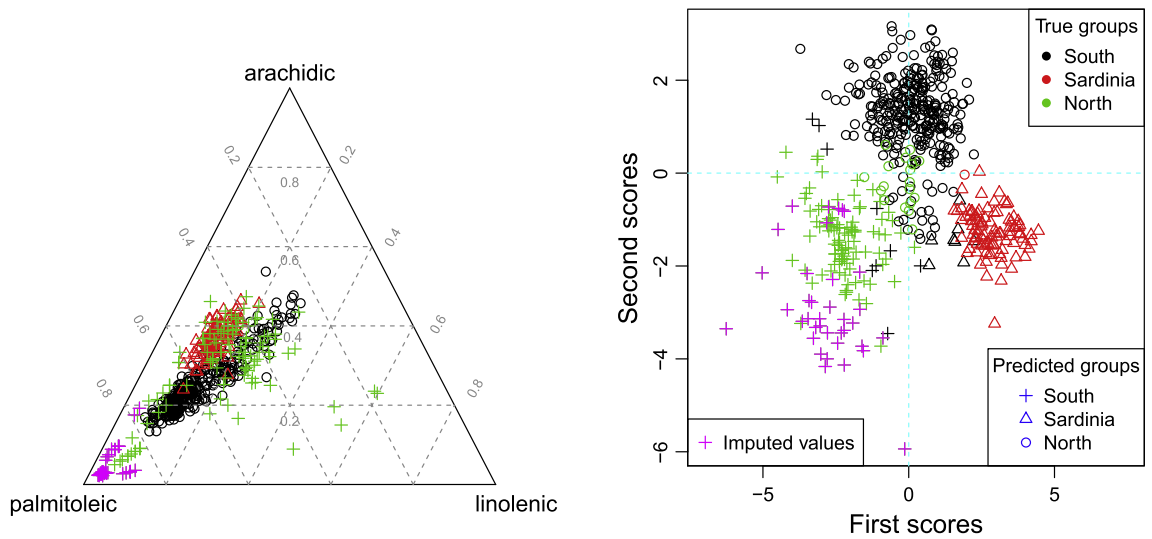

Fig. 10 Ternary diagram (*left*) and result of Fisher discriminant analysis (*right*) of the imputed olive oil data set.

simplex, and the cells that were originally zeros are updated. Next the variable which originally has the second highest amount of zeros is considered, and the same regression procedure as before is applied in pivot coordinates. After each variable containing zero values has been proceeded, one can start the whole process again until the estimated missings stabilize. A detailed description of this algorithm can be found in Templ et al.[49] For the performance of the imputation algorithm it is essential to get an appropriate estimation of the number of latent variables (PLS components) that avoids underfit as well as overfit. For this purpose a bootstrap procedure following Filzmoser et al.[50] is employed. Finally, because the algorithm might get slow with increasing number of components, some numerically more efficient alternatives were proposed, e.g., Chen et al.[51]

Example 6. The Italian olive oil data set which has been used in Example 5 contains zeros for several observations in the compositional parts *linolenic* (30 observations) and *arachidic* (26 observations). All the zeros are from observations of the group of the northern Italian samples, or more precisely, from samples of the sub-regions East Liguria and West Liguria. Accordingly, the model-based imputation as outlined before is applied to the observations of the sub-regions, by using as detection limit half of the minimum non-zero value of the corresponding variable. The ternary diagram in Fig. 10 (*left*) shows the two problematic parts *linolenic* and *arachidic* together with a third part *palmitoleic*. The colors and symbols refer to the three regions; for the imputed values, pink color is used. Indeed, these values are very close to one corner of the triangle, since many samples have very low concentrations for both fatty acids *linolenic* and *arachidic*.

With the complete imputed data set, Fisher discriminant analysis is carried out, and the resulting scores are shown in Fig. 10 (*right*). Again, pink color is used for the imputed samples. They seem to deviate a bit from the whole group of northern Italian oils, which could also be a "true" difference reflected by different sub-regions of origin. All class predictions for these imputed values are correct.

References

1. Aitchison, J. *The Statistical Analysis of Compositional Data*, Chapman & Hall: London, 1986 (Reprinted in 2003 with Additional Material by the Blackburn Press).
2. Egozcue, J. J. Reply to "on the Harker Variation Diagrams;…" by J.A. Cortés. *Math. Geosci.* **2009,** *41* (7), 829–834.
3. Greenacre, M. Variable Selection in Compositional Data Analysis Using Pairwise Logratios. *Math. Geosci.* **2018.** https://doi.org/10.1007/s11004-018-9754-x.
4. Pawlowsky-Glahn, V.; Egozcue, J. J.; Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*, Wiley: Chichester, 2015.
5. Scealy, J. L.; Welsh, A. H. Robust Principal Component Analysis for Power Transformed Compositional Data. *J. Am. Stat. Assoc.* **2015,** *110* (509), 136–148.
6. Filzmoser, P.; Hron, K.; Templ, M. *Applied Compositional Data Analysis*, Springer: Cham, 2018.
7. Eaton, M. L. *Multivariate Statistics: A Vector Space Approach*, John Wiley & Sons: New York, 1983.
8. Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., Eds.; *Compositional Data Analysis in the Geosciences: From Theory to Practice*, Geological Society: London, 2006.
9. Pawlowsky-Glahn, V., Buccianti, A., Eds.; *Compositional Data Analysis: Theory and Applications*, Wiley: Chichester, 2011.
10. Greenacre, M. *Compositional Data Analysis in Practice*, CRC Press: Boca Raton, 2018.
11. R Development Core Team *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria, 2018. http://www.R-project.org.
12. van den Boogaart, K. G.; Tolosana-Delgado, R. *Analyzing Compositional Data with R*, Springer: Heidelberg, 2013.
13. Templ, M.; Hron, K.; Filzmoser, P. robCompositions: An R-package for robust statistical analysis of compositional data. In *Compositional Data Analysis: Theory and Applications;* Pawlowsky-Glahn, V., Buccianti, A., Eds., Wiley: Chichester, 2011; pp 341–355.
14. Palarea-Albaladejo, J.; Martín-Fernández, J. A. zCompositions R Package for Multivariate Imputation of Left-Censored Data under a Compositional Approach. *Chemom. Intel. Lab. Syst.* **2015,** *143*, 85–96.
15. Hron, K.; Filzmoser, P.; Thompson, K. Linear Regression with Compositional Explanatory Variables. *J. Appl. Stat.* **2012,** *39* (5), 1115–1128.

16. Quinn, T. P.; Richardson, M. F.; Lovell, D.; Crowley, T. M. Propr: An R-Package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci. Rep.* **2017,** *7*, 16252.
17. Comas-Cufí, M.; Thió-Henestrosa, S. CoDaPack 2.0: A Stand-alone, Multiplatform Compositional Software. In *CoDaWork'11: 4th International Workshop on Compositional Data Analysis;* Egozcue, J. J., Tolosana-Delgado, R., Ortego, M. I., Eds., Sant Feliu de Guíxols, 2011. ISBN 978-84-87867-76-7.
18. Egozcue, J. J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barcelói-Vidal, C. Isometric Logratio Transformations for Compositional Data Analysis. *Math. Geol.* **2003,** *35* (3), 279–300.
19. Fišerová, E.; Hron, K. On Interpretation of Orthonormal Coordinates for Compositional Data. *Math. Geosci.* **2011,** *43* (4), 455–468.
20. Kynčlová, P.; Filzmoser, P.; Hron, K. Correlation between Compositional Parts Based on Symmetric Balances. *Math. Geosci.* **2017,** *49* (6), 777–796.
21. Egozcue, J. J.; Pawlowsky-Glahn, V. Groups of Parts and their Balances in Compositional Data Analysis. *Math. Geol.* **2005,** *37* (7), 795–828.
22. Pawlowsky-Glahn, V.; Egozcue, J. J. Exploring Compositional Data with the CoDa-Dendrogram. *Austrian J. Stat.* **2011,** *40* (1–2), 103–113.
23. Martín-Fernández, J. A.; Pawlowsky-Glahn, V.; Egozcue, R.; Tolosona-Delgado, J. J. Advances in Principal Balances for Compositional Data. *Math. Geosci.* **2018,** *50* (3), 273–298.
24. Mert, C.; Filzmoser, P.; Hron, K. Sparse Principal Balances. *Statist. Model.* **2015,** *15* (2), 159–174.
25. Pawlowsky-Glahn, V.; Egozcue, J. J. Geometric Approach to Statistical Analysis on the Simplex. *Stoch. Env. Res. Risk Assess.* **2001,** *15* (5), 384–398.
26. Mert, C.; Filzmoser, P.; Hron, K. Error Propagation in Compositional Data Analysis: Theoretical and Practical Considerations. *Math. Geosci.* **2016,** *48* (8), 941–961.
27. Filzmoser, P.; Walczak, B. What Can Go Wrong at the Data Normalization Step for Identification of Biomarkers? *J. Chromatogr. A* **2014,** *1362,* 194–205.
28. Egozcue, J. J.; Pawlowsky-Glahn, V. Changing the Reference Measure in the Simplex and its Weighting Effects. *Austrian J. Stat.* **2016,** *45* (4), 25–44.
29. Hron, K.; Filzmoser, P.; de Caritat, P.; Fišerová, E.; Gardlo, A. Weighted Pivot Coordinates for Compositional Data and their Application to Geochemical Mapping. *Math. Geosci.* **2017,** *49* (6), 797–814.
30. Jolliffe, I. T. *Principal Component Analysis,* 2nd edn; Springer: New York, 2002.
31. Kynčlová, P.; Filzmoser, P.; Hron, K. Compositional Biplots Including External Non-compositional Variables. *Statistics* **2016,** *50* (5), 1132–1148.
32. Filzmoser, P.; Hron, K.; Reimann, C. Principal Component Analysis for Compositional Data with Outliers. *Environmetrics* **2009,** *20,* 621–632.
33. Aitchison, J.; Greenacre, M. Biplots of Compositional Data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **2002,** *51* (4), 375–392.
34. Gabriel, K. R. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika* **1971,** *58* (3), 453–467.
35. Rousseeuw, P. J.; Van Driessen, K. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* **1999,** *41* (3), 212–223.
36. Filzmoser, P.; Hron, K. Outlier Detection for Compositional Data Using Robust Methods. *Math. Geosci.* **2008,** *40* (3), 233–248.
37. Filzmoser, P.; Maronna, R.; Werner, M. Outlier Identification in High Dimensions. *Comput. Stat. Data An.* **2008,** *52* (3), 1694–1711.
38. Bruno, F.; Greco, F.; Ventrucci, M. Spatio-Temporal Regression on Compositional Covariates: Modeling Vegetation in a Gypsum Outcrop. *Environ. Ecol. Stat.* **2015,** *22* (3), 445–463.
39. Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis,* 6th edn; Prentice Hall: Upper Saddle River, 2007.
40. Varmuza, K.; Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press: Boca Raton, 2009.
41. Rousseeuw, P. J.; Van Driessen, K. Computing LTS Regression for Large Data Sets. *Estadistica* **2002,** *54,* 163–190.
42. Wold, H.; Sjöström, M.; Eriksson, L. PLS Regression: A Basic Tool of Chemometrics. *Chemom. Intel. Lab. Syst.* **2001,** *58,* 109–130.
43. Filzmoser, P.; Hron, K.; Templ, M. Discriminant Analysis for Compositional Data and Robust Parameter Estimation. *Comput. Stat.* **2012,** *27* (4), 585–604.
44. Kalivodová, A.; Hron, K.; Filzmoser, P.; Najdekr, L.; Janečková, H.; Adam, T. PLS-DA for Compositional Data with Application to Metabolomics. *J. Chemometr.* **2015,** *29* (1), 21–28.
45. Brereton, R. G. *Chemometrics for Pattern Recognition*, Wiley: Chichester, 2009.
46. Walach, J.; Filzmoser, P.; Hron, K. Data Normalization and Scaling: Consequences for the Analysis in Omics Sciences. In *Data Analysis for Omics Sciences: Methods and Applications;* Jaumot, J., Bedia, C., Tauler, R., Eds., Elsevier: Amsterdam, 2018; pp 165–196.
47. Hron, K.; Templ, M.; Filzmoser, P. Imputation of Missing Values for Compositional Data Using Classical and Robust Methods. *Comput. Stat. Data Anal.* **2010,** *54* (12), 3095–3107.
48. Martín-Fernández, J. A.; Barcelaó-Vidal, C.; Pawlowsky-Glahn, V. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Math. Geol.* **2003,** *35* (3), 253–278.
49. Templ, M.; Hron, K.; Filzmoser, P.; Gardlo, A. Imputation of Rounded Zeros for High-Dimensional Compositional Data. *Chemom. Intel. Lab. Syst.* **2016,** *155,* 183–190.
50. Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated Double Cross Validation. *J. Chemometr.* **2009,** *230* (4), 160–171.
51. Chen, J.; Zhang, X.; Hron, K.; Templ, M.; Li, S. Regression Imputation with Q-Mode Clustering for Rounded Zero Replacement in High-Dimensional Compositional Data. *J. Appl. Stat.* **2018,** *45* (11), 2067–2080.