News from "compositions", the R package

M. Bren[1], R. Tolosana-Delgado[2], and K. G. van den Boogaart[3]
[1] University of Maribor, Slovenia, *matevz.bren@fov.uni-mb.si*
[2] University of Göttingen, Germany ,
[3]University of Greifswald, Germany

Abstract

The R-package "compositions" is a tool for advanced compositional analysis. Its basic functionality has seen some conceptual improvement, containing now some facilities to work with and represent ilr bases built from balances, and an elaborated subsystem for dealing with several kinds of irregular data: (rounded or structural) zeroes, incomplete observations and outliers. The general approach to these irregularities is based on subcompositions: for an irregular datum, one can distinguish a "regular" subcomposition (where all parts are actually observed and the datum behaves typically) and a "problematic" subcomposition (with those unobserved, zero or rounded parts, or else where the datum shows an erratic or atypical behaviour). Systematic classification schemes are proposed for both outliers and missing values (including zeros) focusing on the nature of irregularities in the datum subcomposition(s).

To compute statistics with values missing at random and structural zeros, a projection approach is implemented: a given datum contributes to the estimation of the desired parameters only on the subcompositon where it was observed. For data sets with values below the detection limit, two different approaches are provided: the well-known imputation technique, and also the projection approach.

To compute statistics in the presence of outliers, robust statistics are adapted to the characteristics of compositional data, based on the minimum covariance determinant approach. The outlier classification is based on four different models of outlier occurrence and Monte-Carlo-based tests for their characterization. Furthermore the package provides special plots helping to understand the nature of outliers in the dataset.

Keywords: coda-dendrogram, lost values, MAR, missing data, MCD estimator, robustness, rounded zeros.

1 Introduction

The R-package "compositions" (van den Boogaart and Tolosana-Delgado, 2008) is a tool for advanced compositional analysis. The package provides several layers: compositional arithmetic (e.g. perturbation, centering), estimation of distributional parameters (e.g. compositional mean, variation matrix), statistical graphics (e.g. multiple ternary diagrams), multivariate analysis (e.g. principal components) and statistical tests (e.g. linear models) for four different types of closed and unclosed compositions. The basic functionality and philosophy of the package is encapsulated in a few simple commands. This paper presents the recent developments of this package, directed towards the analysis of irregular compositional data, and to facilitate working with ad-hoc ilr bases.

Irregular data in compositions are quite common, and present an important challenge due to the inherent and inextricable multivariate nature of these kind of data. While in real data sets one can define several kinds of irregularities by marginals, like a variable completely missed at random, or an atypical marginal value, such a univariate approach is misleading in compositional data, just due to the binding effect of the closure.

Based on the vector space approach for the simplex we implemented a new strategy to treat atypical observations, zeros and missing values in the R package "compositions". The main idea is based on the realization that *regular* (i.e. fully observed and with a typical behavior) parts define a regular subspace where inference is possible with classical tools, whereas the unobserved or atypical variables specify some (mostly, qualitative) information on the orthogonal complement of the regular subspace.

With this approach, we can introduce missing values, either (a) structural zeros, (b) missing values at random, or (b) values below the detection limit, into graphics, descriptive statistics, statistical computation and statistical models for compositional data. This approach allows a consistent handling, without the usual problems introduced by imputation techniques and propagated to the "sane" parts by the closure. The idea for data with missing (in a general sense) variables is to use each datum only on the subcomposition of its fully observed parts. However, quick imputation techniques are also provided within the package, for the sake of comparison.

On the other side, compositional outliers are individuals with an atypical composition in any of its parts, any of its log-ratios or any relationship between log-ratios, as outliers in conventional multivariate analysis may have an atypical measured variable or an atypical relationship between variables (Rousseeuw and Leroy, 2003). The extra difficulty of outliers in compositional data lies on the fact that a single wrong measurement might distort the portions of all parts. It is therefore difficult to identify the source of error with classical outlier detection techniques. However, log-ratios of "sane" parts are not modified. Based on this idea we propose a classification for compositional outliers according to different possible explanations, like: (a) single component measurement errors, (b) equilibrium distortion, (c) population mixture and contamination, or (d) atypical subpopulations. Each of these explanatory hypothesis and each possible subcomposition define a particular affine subspace of the simplex, going through the mean of the population, where the datum is typical. Then, every individual datum can be classified according to its Mahalanobis distance to each of these subspaces, computed using robust estimates of the mean and covariance matrix. This approach leads to a hierarchy of hypothesis on the origin of the distortion, which can be intuitively described as an ordering by "proximity to typicality". This classification concept is implemented in several graphical tools and tests, to explore the probable character of the outliers of a dataset.

All these missings and outlier types will be illustrated throughout the text by using some simulated 3-part compositions, easy to plot and understand. The R generating functions are included in appendix A. One may think that the presented techniques are not necessary in 3-part compositions, because we can easily spot irregular data in ternary diagrams. However, the proposed methods are indispensable in higher dimensional compositions, where visual inspection fails.

2 Balance bases and coda-dendrogram

Recall that the compositional operations of perturbation, powering (respectively denoted as $\oplus$ and $\odot$), and the scalar product

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{D} \sum_{i>j}^{D} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j} = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) \rangle_{euc}, \tag{1}$$

defined by Aitchison (1986) build an Euclidean space of the simplex (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001), with the clr or centered log-ratio transformation (Aitchison, 1986) defined as

$$\mathrm{clr}_i(\mathbf{x}) = \log x_i - \frac{1}{D} \sum_{j=1}^{D} \log x_j,$$

and inverse $\mathbf{x} = \mathcal{C}[\exp(\mathrm{clr}(\mathbf{x}))]$ (exponent applied by components). In these expressions and in the following developments, D denotes the number of parts, and $\mathcal{C}[\cdot]$ the closure operation. Actually, the clr is an isometry between the simplex and a $D-1$ dimensional hyperplane of $\mathbb{R}^D$ (the real space) orthogonal to the vector $\mathbf{1} = [1, \ldots, 1]$, the bisector of the first orthant. This plane may be called the clr-plane.

Like all vectors from any Euclidean space, a composition can be univocally identified with a set of $(D-1)$ coordinates with respect to a basis. In general, to ease proofs and computations, it is useful to work with coordinates with respect to orthonormal bases (Pawlowsky-Glahn, 2003), for compositions computed with the isometric log-ratio transformation (ilr, Egozcue et al., 2003),

$$\mathrm{ilr}(\mathbf{x}) := \ln(\mathbf{x}) \cdot \mathbf{V} = \mathrm{clr}(\mathbf{x}) \cdot \mathbf{V}, \quad \text{with} \quad \mathbf{V} \cdot \mathbf{V}^t = \mathbf{I}_{D-1} \quad \text{and} \quad \mathbf{V}^t \cdot \mathbf{V} = \mathbf{I}_D + \alpha \cdot \mathbf{1}_{D \times D},$$

for $\mathbf{I}_D$ and $\mathbf{1}_{D \times D}$ respectively the identity matrix and the matrix full of ones, each of the indicated dimensions, and any scalar α ($= 1/D$, typically). But ilr bases may be quite difficult to interpret, as often each coordinate involves several parts. It has been thus recommended to go the other way round: work with coordinates with respect to an interpretable, orthonormal basis. Sequential binary partitions are the way to build such ad-hoc bases (Egozcue and Pawlowsky-Glahn, 2005, 2006).

In a sequential binary partition, the parts of a composition are divided in two groups, and each subgroup is further split in two, in a series of sequential steps, until all "groups" have only one part. Equivalently, one can follow the opposite strategy, grouping parts by pairs like in a hyerarchical cluster analysis. Such partition structure can be written in a matrix with $D-1$ columns (one for each partition step) and D rows (one for each part), full of $-1, 0$ or $+1$ values, called in "`compositions`" a `signary` matrix. With each column of the signary matrix, one can obtain a vector of the basis as follows:

- if we have p times $+1$ and n times -1 in a column, scale them respectively to $+n$ and $-p$, so that now the column sums up to zero,
- then divide these numbers by their euclidean norm, $\sqrt{p \cdot n^2 + n \cdot p^2}$.

These steps have been implemented in the function `gsi.buildilrBase`, taking a signary matrix encoding a binary partition, and returning a $\mathbf{V}$ matrix.

To define a sequential binary partition, one can follow criteria of reasonable association or interpretability, without a look at the data. Or one can apply a variable clustering technique and derive the partition structure from the resulting tree. This has been implemented in the function `gsi.merge2signary`, which takes the tree structure of the output of a cluster analysis (element `merge` from the result of an `hclust(...)` call) and returns a signary matrix (which can be afterwards fed to `gsi.buildilrBase`).

Two applications of this clustering idea might be useful: looking for groups of parts with more similar *compositional* patterns, or looking for groups of parts with more similar patterns of *missing values*. The steps would be:

1. define the matrix of distances between the parts; in the first case, one can use the variation matrix; in the second case, the Euclidean distance between the indicator-transformed columns (putting 0 if the datum was observed and 1 if missing, or vice versa)
2. apply cluster analysis with the desired clustering method
3. transform the merging structure into a signary matrix,
4. and transform the signary matrix into a matrix of ilr definition,

```
> X = acomp(X)
> dd = as.dist(variation(X))
> hc = hclust(dd, method = "ward")
> w = gsi.merge2signary(hc[["merge"]])
> V = gsi.buildilrBase(w)
```

This sequence would yield a basis following the compositional similarity criterion. The missingness similarity criterion is quite standard, and it has been already implemented in the basic function of ilr basis computation: `ilrBase(D,method="optimal")` produces the ilr basis for a D-dimensional composition with less influence of the missing values, i.e. where the number of coordinates of all observations that we can compute is the highest.

To end this section on the ilr functionalities within "`compositions`", the package has also available the CoDa-dendrogram (Egozcue and Pawlowsky-Glahn, 2005). A CoDa-dendrogram is a graphical representation of the tree structure explained before, jointly with some summary, one-dimensional representations of each of the empirical marginal distributions of the coordinates. For instance, using the hydrochemical data set from the Llobregat River basin included in "`compositions`" and the basis defined with the variation array, this

```
> CoDaDendrogram(X, V, type = "boxplot", range = NULL)
```

would generate the dendrogram presented in Figure 1. This function has a lot of optional arguments, explained in its help file (`?CoDaDendrogram`). The most important are those encoding the basis (either `V`, `signary` or `mergetree`, accepting an ilr basis, a signary partition encoding or a cluster merge structure), and the `type` of output to be plotted in each ilr segment: it can draw boxplots (as in the example, the default), rugs of the data (`="points"`), only the vertical lines displaying the variance of that ilr coordinate (`="lines"`), and histograms (`="histogram"`) and kernel density estimation (`="density"`).

3 Classification of irregular data

3.1 The types of missing values

There are several types of "missing values" and "zeros" typically present in compositional data:

1. *Below detection limit* (BDL): as its name says, this variable is lost because the detection limit of the analytical technique was higher than the actual value of the element. In classical statistics, this is a censored value. But in the relative geometry of Aitchison, it may be to some extent seen as a missing, since a BDL can give rise to extremely different ratios: even a clever

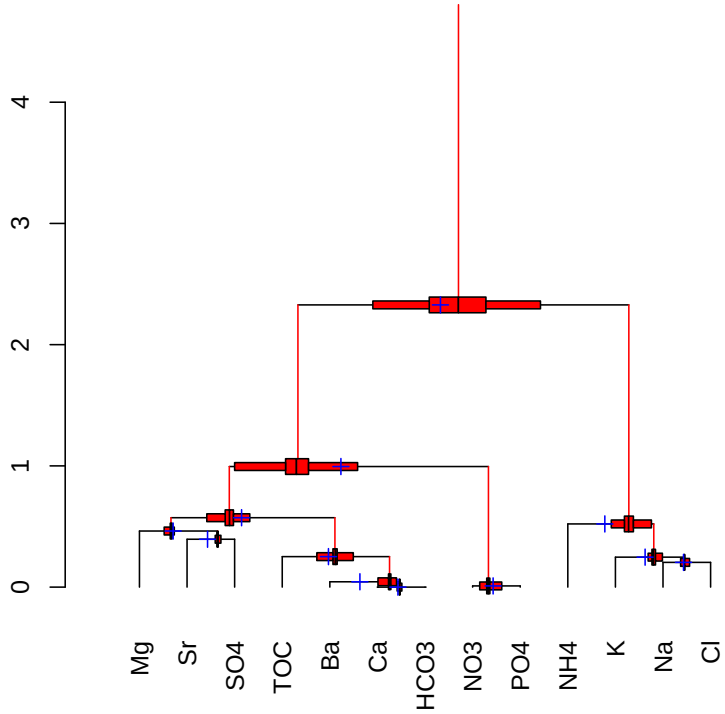

Figure 1: CoDaDendrogram of the Hydrochemical data set, based on the criteria of observed association through the variation matrix. Each box plot extends over the range (-6.67;8.81), and a blue cross has been added to mark the origin.

imputation replacing of the values will yield very different ratios depending on the rest of the dataset, while drastically reducing the variance of those log-ratios where the replaced parts are involved. This is one of the reasons for the artefacts observed with imputation and the major source of all criticism on the Aitchison approach to the simplex. EM-based imputation techniques have been recently developed to deal with this case (Palarea-Albadalejo et al., 2007; Palarea-Albadalejo and Martín-Fernández, 2008), with promising results, but they are not yet available in "`compositions`".

2. *Structural Zero* (SZ): in some cases it makes no sense to measure (or simply talk about) a specific component, e.g. water content in dried out material, since the specific component has been removed or *cannot* be there. That is a structural zero. One could argue (Aitchison, 1986) that in this case we have two different population. However if the structural zero is unrelated to the process to be investigated, the relative geometry of the simplex is well able to understand that the observed value is just a subcomposition of the original one.

3. *Missing at random* (MAR): the measurement process just failed at random, without any relation to the actual lost value, thus we have no data for a point. In compositions, this happens to be at first sight very near to the structural zero, even though the understanding is very different: for MAR, one says that there exists a true but unkown value, while SZ neglects the possibility of talking about it. Note that the distinction of "completely" missing at random (being based on whether the missing process depends on the observed variables or not, while not depending on the lost variable itself) is very complex in compositional data,

due to the binding effect of the closure.

4. *Missing not at random* (MNAR): the value was lost, but the missing process is intimately related to the lost value itself. This is a problematic case in all statistics. It could arise e.g. from measurement procedures with a different probability to fail or not being performed stochastically dependent to their actual value. Theoretically an analysis based on such data is only possible with a clear model for the missing process invovled. However, from a practical point of view, since one can hardly distinguish MAR and MNAR from the data set alone, we will analyse the data in a MAR fashion, and the analyst should bear this problem in mind.
5. *Not missing value* (NMV): to allow an easy reference within the package, we added the abbreviation NMV to label those observed, not missing values.

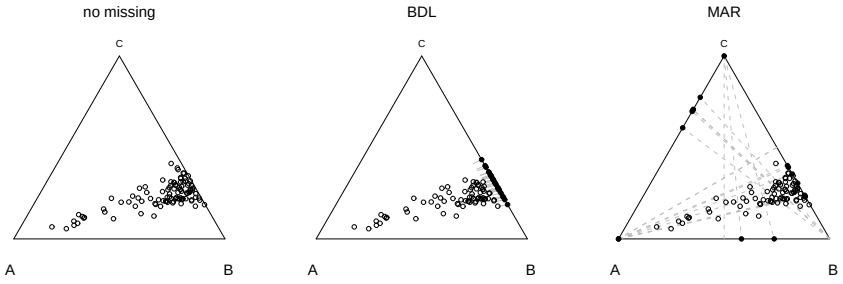

Figure 2: Ternary diagrams of example datasets with missing values. Empty circles represent fully-observed data, whereas filled circles on the edges correspond to values with a lost component. For these last cases, the actual value is somewhere on the grey segment.

Figure 2 shows three synthetic data sets, each generated following the missing process corresponding to its code:

dataNMV is a random generation with normal distribution in the simplex, without further modification;

dataBDL is the preceding data set where all values below 5% have been replaced by zero;

dataMAR is the dataNMV basic data set, where randomly 5% of values have been removed.

For all these cases, grey segments have been plotted representing the possible range of actual values of each compositional datum with a missing part. Note that these segments are iso-proportion lines, with a fixed proportion of the two non-missing parts corresponding to their actually-osberved ratio. In the case of MAR values, any composition with that particular ratio is possible, thus the segments are in fact whole Aitchison lines. In the case of BDL, the segments are Aitchison semi-lines from the composition with lost value equal to the detection limit (5% in this case).

We do not include MNAR, because to characterize them we would need to specify a missing process, which is by no means something general. Structural zeroes are also not illustrated with a specific example, being similar to MAR in the absence of relationship between the zero-affected part and the rest of the data set. If a MAR representation of structural zeroes would not homogeneously intersect the body of fully-observed data, we could suspect that the SZ have an influence and proceed splitting the sample, as recommended by Aitchison (1986).

3.2 Conceptual types of outliers

To provide tools to distinguish between the possible types of outlier, we need to understand what is their characteristic atypicality.

1. *Extreme observations*: improbable randomly generated from the typical composition. They should follow the probability law, e.g. plot adequately on a QQ plot with confidence bands.

2. *Measurement errors*: these are errors in some individual portions. (e.g. instrumental errors or typos). Once the affected part(s) is removed, the remaining subcomposition of such an individual should nicely fall within the typical portion of the sample.

3. *Individual atypical samples*: these are "lonely" samples which do not fit in the typical distribution, but do not have any specific behaviour (e.g. erratic samples, or a mistake during the sampling campaign).

4. *Non-normality*: if the actual distribution is not normal (e.g., skewed or heavy tailed, as could happen with a distorted equilibrium) a significant portion of the sample will not be classified as typical; these could be confirmed with distribution tests, but it is enough to find many extreme values (of the first type) farther than a typical outlier limit from the mean (or confined in a half-space, in general) to suspect for non-normality of the distribution.

5. *Contaminated samples*: these show a behaviour similar to the preceding one, but they should more or less follow a mixture segment from the typical portion of the data to the composition of the contaminant, and be bounded between these two end-members. Whether this segment is compositional or "straight" will depend on the contamination mechanism.

6. *Multimodality*: the existence of subpopulations might yield a bi- or eventually multi-modal distribution (e.g. originating from several rather different facies/groups in the observed region, or just due to replacement of a common BDL zero by a fraction of the detection limit); we expect to detect such secondary modes as clusters of atypical values.

Figure 3 shows six synthetic data sets, each generated following the process corresponding to its number:

data1 is a random generation with normal distribution in the simplex without specific distortions; it can eventually have extreme values, but no atypical value; this is the base upon which all other data sets are built:

data2 has approximately 10% of individuals affected with a random additive distortion in the first component;

data3 has a single contrasted individual composition $[A, B, C] = \mathcal{C}[0.5, 1.5, 2]$;

data4 follows a Cauchy distribution in the simplex, defined as the inverse ilr transformation of a multivariate Cauchy distribution in a 2-dimensional space; the Cauchy distribution is an example of an extreme heavy-tailed distribution;

data5 has approximately 10% of individuals contaminated as a (random, additive) mixture with the composition $[A, B, C] = \mathcal{C}[0.1, 1, 2]$;

data6 has 16% of individuals in a subpopulation with the same variance structure and a mean on $[A, B, C] = \mathcal{C}[4, 4, 1]$.

4 Inference in the presence of irregular data

This section briefly introduces the theoretical concepts behind our approach. It is not intended as a rigorous exposition, but included here for the sake of completion. Proper proofs, developments and arguments will be published in brief.

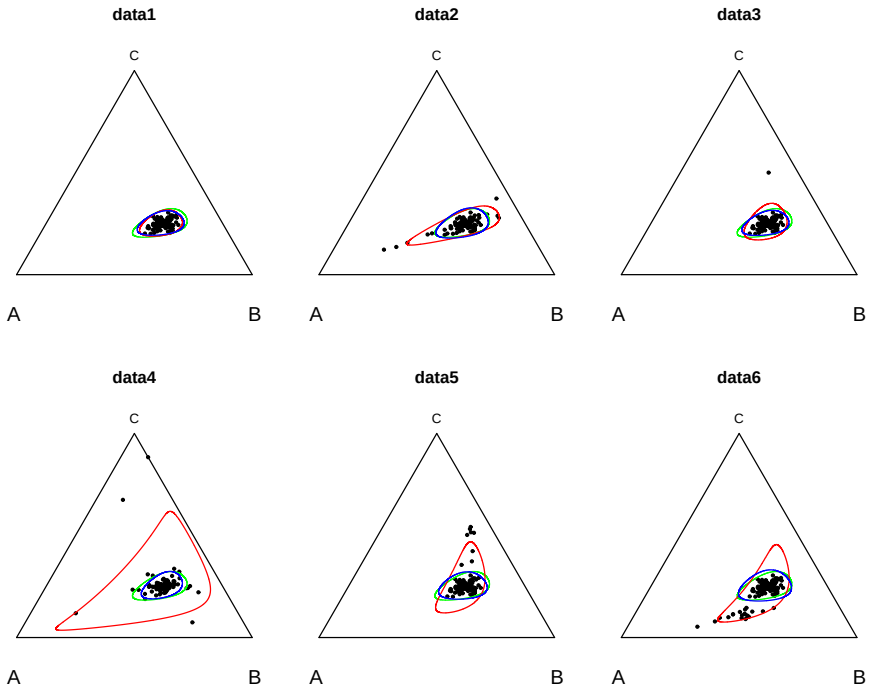

Figure 3: Ternary diagrams of example datasets with outliers. Ellipses represent the true 3σ area of the non-outliers (blue), and the correspondig 3σ ellipses for the estimated covariance matrix and average, using the minimal covariance determinant estimator with maximum breakpoint (green, see section 4.1 for details), and the non-robust classical method (red). The ellipses of this last case are strongly different from the other two.

4.1 The minimum covariance determinant estimator for compositions

The concepts of robust statistics, and even robustness itself, have a strong geometric grounding. A statistic is thought to be non-robust if it can change arbitrarily far from its original value by *translating* a single observation away from the rest (Huber , 1981). And the common definition of robust mean and robust covariance matrix makes use of the parallelism between these two statistics and an *ellipse*: with a center $\mathbf{m}$ (a *mean*) and a positive-definite form $\mathbf{T}$ (the *inverse* of a variance matrix) we can define one and only one ellipse as the set of points $\mathbf{x}$ fulfilling:

$$\langle(\mathbf{x}-\mathbf{m})\cdot\mathbf{T},\mathbf{x}-\mathbf{m}\rangle_{euc} = (\mathbf{x}-\mathbf{m})\cdot\mathbf{T}\cdot(\mathbf{x}-\mathbf{m})^t = 1.$$

Rawly speaking, the minimum covariance determinant technique looks for the smallest ellipse (in the sense of lowest determinant) containing $100(1-\alpha)\%$ of the data (with breakpoint $\alpha \leq 0.5$), takes its center as robust estimate of the mean and its positive definite matrix as robust estimate of the covariance matrix, inverted and conventiently scaled to account for the α level used (Rousseeuw and Leroy, 2003).

For compositions, this technique has been applied to the alr (Aitchison, 1986), clr and ilr (Egozcue et al., 2003) transformed data, with equivalent results (Filzmoser and Hron, 2008). In this last case, one must take care to adequately invert the clr-covariance matrix by using Moore-Penrose generalized inverses, and computes its determinant as the product of its first $D-1$ eigenvalues (van den Boogaart et al., in prep).

4.2 The projection-estimation approach

Recall that the clr transformation is an isometric mapping from the simplex to the clr plane. This hyperplane is spanned by vectors $\mathbf{l}_i = \mathbf{e}_i - \frac{1}{D}\mathbf{1}$ one for each component (which do not form a basis because they are not linearly independent), with $\{\mathbf{e}_i\}$ the canonical basis of $\mathbb{R}^D$. For example, $\mathbf{1}_2 = [-1, D-1, -1, \ldots, -1]/D$.

One of the possible characterizations of compositions given by Aitchison (1986) was the full log-ratio vector, the set of all its pairwise log-ratios. If we have a missing value in a component, then we cannot compute those log-ratios involving that part, thus we are only informed about the complementary subcomposition. However, from Aitchison (1986, p. 34) one can deduce that the operation of building a subcomposition can be seen as a projection P_M of the clr values into the subspace of the clr-plane orthogonal to all $\mathbf{l}_i, i \in M$ of the non-selected components $(x_i)_{i\in M}$. Note that here M is the set of non-selected parts, which might be a counter-intuitive notation for subcomposition building, but will become natural when we identify the non-selected parts with those parts missing.

Two examples (with $D = 5$ parts) might help in illustrating what P_M is. Take $M = \emptyset$ empty, i.e. all parts were observed: in this case, $P_\emptyset$ projects a datum from the clr plane onto the clr plane:

$$P_\emptyset \mathrm{clr}(\mathbf{x}) = \underbrace{\frac{1}{5}\begin{pmatrix} 4 & -1 & -1 & -1 & -1 \\ -1 & 4 & -1 & -1 & -1 \\ -1 & -1 & 4 & -1 & -1 \\ -1 & -1 & -1 & 4 & -1 \\ -1 & -1 & -1 & -1 & 4 \end{pmatrix}}_{P_\emptyset} \cdot \mathrm{clr}(\mathbf{x}) = P_\emptyset \cdot \log(\mathbf{x}),$$

i.e. the matrix allowing to compute the 5-part clr from the log-transformed 5-part composition, and acting as identity within the clr plane. Take now $M = \{2, 5\}$, i.e. we observed the subcomposition $\{1, 3, 4\}$, thus $P_{\{2,5\}}$ projects a datum from the clr plane onto the subspace orthogonal to $\mathbf{1}_2 = [-1, 4, -1, -1, -1]/5$ and $\mathbf{1}_2 = [-1, -1, -1, -1, 4]/5$, namely

$$P_{\{2,5\}} \mathrm{clr}(\mathbf{x}) = \underbrace{\frac{1}{3}\begin{pmatrix} 2 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 2 & -1 & 0 \\ -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{P_{\{2,5\}}} \cdot \mathrm{clr}(\mathbf{x}) = P_{\{2,5\}} \cdot \log(\mathbf{x}),$$

which is now the matrix allowing to compute the 3-part clr from a log-transformed 3-part composition, but conveniently placed on a 5×5 matrix, with those missing parts removed from the computation by the expeditious mechanism of multiplying them by 0, and resulting always in an image of 0.

The idea of the projection-estimation approach is to represent the information we get from a compositional datum with missings by $P_M \log(\mathbf{x})$ and P_M itself, where M is the set of missing parts on *that specific datum*. Then, since each datum $\mathbf{x}_i$ has a different set of missing parts M_i, each datum has a different associated projector P_{M_i}. An unbiased estimator $\hat{\mathbf{m}}$ of the mean with missings can thus be obtained by:

$$\mathrm{clr}(\hat{\mathbf{m}}) := \left(\sum_i^N P_{M_i}\right)^- \sum_i^N P_{M_i} \log(\mathbf{x}_i) \tag{2}$$

where A^- denotes the Moore-Penrose-Inverse of A, and N is the number of data. In this way only observed parts of every datum contribute to the mean.

The obtention of an unbiased estimator of the variance in the presence of compositional missings is quite tricky,

$$c(\hat{var}(\mathrm{clr}(\mathbf{x}))) := \left(\sum_{i\neq j}^{N} P_{M_i\cap M_j}\otimes P_{M_i\cap M_j}\right)^{-}\times \\ \times\sum_{i\neq j}^{N} c\left(P_{M_i\cap M_j}(\log(\mathbf{x}_i)-\log(\mathbf{x}_j))(\log(\mathbf{x}_i)-\log(\mathbf{x}_j))^t P_{M_i\cap M_j}\right)$$

involving Kronecker products ($\otimes$) of the projectors and the conversion $c(\cdot)$ of a $D\times D$-matrix in a vector of D^2 components. A proper proof can be found in van den Boogaart et al. (2006), but the idea is to work with the variation matrix (the set of all pair-wise log-ratios, obtained with the terms $(\log(\mathbf{x}_i)-\log(\mathbf{x}_j))(\log(\mathbf{x}_i)-\log(\mathbf{x}_j)$, and then project it using the projectors of both datum i and j (thus, obtain only those pairs which are both times observed). The averaging is finally done as with the mean, downweighting the contribution of those pairs with more missing values.

5 Dealing with missing values and zeroes

5.1 Replacement and coordinates

The common method of treatment of missing parts is to replace them with a reasonable number. The command `zeroreplace(x,d,a=2/3)` does it: `x` is the composition with missings, `d` is the vector of detection limit, and `a` is the fraction of this limit to be used for replacement. Note that this function is intended for BDL missings, as in MAR values one should rather prefer a replacement by "the mean", and in SZ any replacement is simply meaningless.

Replacement is often seen as operating on the original, compositional values. But what happens with coordinates? Given that any coordinate is the quotient of at least a pair of parts, it is illustrative to think which information do we have about the log-ratio of two parts, depending on their classes of missing/non-missing.

- The log-ratio of two observed values (NMVs) does not entail any problem: it can be computed without further ado. Coordinates involving only NMVs will be fully computable, and will be plottable as points.
- The log-ratio of a MAR or a SZ with a value of any kind can have respectively *any* value or no value at all. Coordinates involving such a log-ratio can have virtually any value, thus they give the representation of a datum one more degree of freedom. We will be able to represent MAR values as lines. Log-ratios of MAR and NMVs can be replaced by the average of that log-ratio, as they are still MAR values.
- The log-ratio of a BDL with a NMV is unknown but has a bound: if the missing part is $x_i < d$ and the observed one is x_j, then we know that the true value $\log(x_i/x_j) < \log(d/x_j)$. Coordinates involving NMVs and BDLs (*the latter on the same part of the quotient*!) are thus censored values, to be replaced by a sensible value below the bound.
- The log-ratio of two BDLs can have virtually *any* value, thus behaving quite like a MAR. Coordinates with BDLs in the numerator and the denominator cannot be determined, nor bounded, and may be quite safely replaced by any suitable value, e.g. the mean.

The use of the optimal base provided by `ilrBase(D,method="optimal")` explained in section 2 not only maximizes the amount of coordinate values that can be computed, but also offers a structured way to replace values attending to these "rules" of propagation of the class of missingness from the composition to its coordinates.

5.2 Representing missing values in scatter plots

To represent compositional data sets with values missing at random, we have already used segments covering the set of points in the ternary diagram compatible with the available information (Fig. 2). Note that in ternary diagrams, each datum with a missing value is plotted as an isoproportion line, i.e. a line passing through the vertex of the missing component and having a constant ratio of the two observed components. If two parts are missing, the datum cannot be represented in that ternary diagram.

The same can be done if one wants to represent a set of log-ratios (e.g. ilr or alr coordinates), like in Figure 4. Note that in this simple 3-part case, the lines compatible with lost values can only be at 0^o (for missings in A), 60^o (in C) or 120^o (in B) from the X axis (with alr they would be at 0^o, 45^o and 90^o respectively). If we had more parts these angles may be completely different.

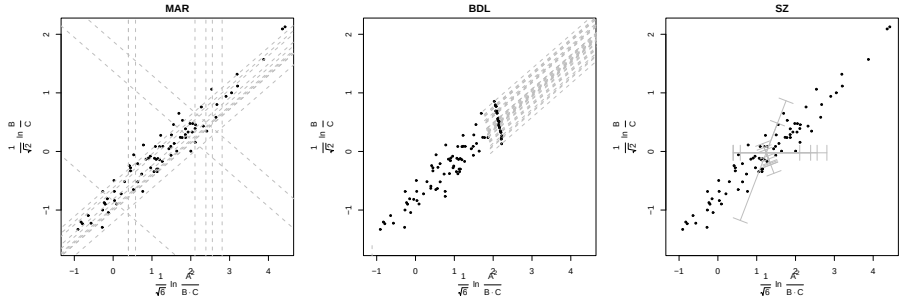

Figure 4: Coordinate representation of the data sets with missings, following the same idea than with ternary diagrams: grey lines show the set of points compatible with each observation with missing variables. A representation of SZ is also proposed (using MAR data), where the grey lines represent rather the data subsets with a missing value as 1-dimensional data sets on axes placed on the average.

Note that this representation of missing values as straight lines might help in deciding whether a set of missing values are MAR, SZ or BDL. For MAR (or SZ without influence on the observed part), the set of lines should be homogeneously covering the whole domain, without preferentially covering any area. In contrast, for SZ with influence on the composition, they should present a pattern. If this pattern is a clear-cut censoring of the lower values, then we may consider that the data set is affected of BDL missingness.

In this line, if one wants to distinguish values which are positively know to be SZ from other MAR, an option might be to use rugs instead of lines. In the MAR representation, we plot all possible points compatible with the observed subcomposition; in the SZ representation, an axis is drawn through the average of the data set, and on each axis we place a marker at the observed log-ratio value. This representation does not visually transmit the idea that there is an actual value that *might* be somewhere, while still allowing comparison of the fully-observed data set against the data subsets with SZ.

Values below the detection limit have been represented with semi-lines, because in this case we know that the value cannot be above a certain threshold. In Figure 4 the same representation is adopted for log-ratio scatterplots.

In 2D-representation of three-part compositions each datum is either lost or observed. But in higher dimensions, an interesting third situation might occur: a particular datum *can* be represented in a given ternary diagram or log-ratio scaterplot, *but* it has missing values in a non-represented part – in these cases, it is advisable to highlight those points, specially to see whether they are MAR or SZ.

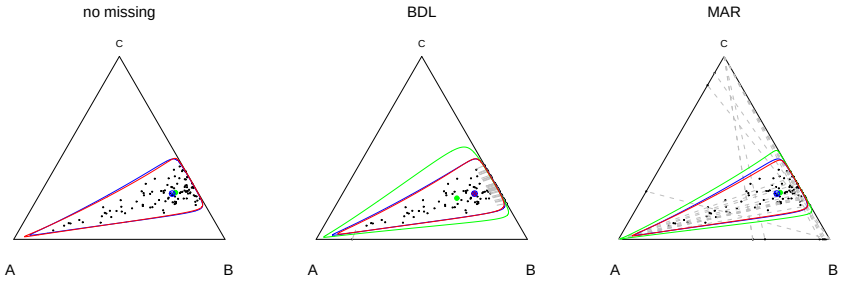

Figure 5: Example data sets with missing values, together with several mean and variance estimates. Blue represents the true population parameters, red the classical estimates (when needed, with replacement), and green the estimates provided by the projection-estimation approach. The means are represented by dots/circles, and ellipses show 3σ areas around them.

5.3 A comparison of projection-estimation and replacement statistics

Figure 5 displays as a dot and an ellipse the estimates of mean and covariance matrix obtained with the projection-estimation approach, compared with the population values and the estimates delivered by a replacement technique, using 2/3 of the detection limit of 5%. For the MAR type, one should not replace the lost values using that limit, but using an average: for instance the mean estimate provided by the projection-estimation technique (Eq. 2). A quick look at the figure shows that the projection-estimation technique delivers good results for the MAR case, though the variance is slightly overestimated, as is expected given that it suffers of lack of information. For the BDL case, the projection-estimation technique overestimates the proportions of the lost parts, as this technique completely ignores the lost data. On the contrary, the replacement technique delivers better results for BDL cases, because it takes into account the fact that the lost values are below the detection limit. With this realization, we see that the projection-estimation technique overestimates the proportions of those parts with BDL values (both in the mean and the spread), as the estimated mean is displaced towards the vertices A and C and the variance ellipse is also much larger towards these. Finally, if we "plug" the projection-estimation mean on the estimation of the variance by the replacement technique in the MAR case, the results are excellent.

6 Dealing with outliers

6.1 Detection of individual outliers

In section 3.2 we listed six characteristic atypicalityes of outliers. With these one can establish a hierarchy for hypothesis 1 to 6, as each one is quite improbable if compared with its preceding one as an alternative hypothesis. For instance, outliers of types 3 will very rarely become typical after removing a single part, as happens with type 2 generated by a single wrong measurement. Thus we can not positively decide for type 2 but exclude it, e.g. if several measured parts need to be ignored to make the composition typical. In a similar way, to decide for a non-normal distribution (type 4 or 6) it is not enough with a couple of atypical values: one needs a bunch of them, outstretched over a large range of values (which might go in one direction or a halfspace from the center of the data if the distribution is skew) for type 4, or tightly clustered together for type 6.

We are aware of the fact that this is not a complete list of possible causes for atypicality, but only those considered in the further developments. However, since we provide visualization tools of this classification, the analyst can see if a given sample or group of samples does not satisfy any of these patterns, and look then for a possible alternative cause. Most of these types of outliers

can also be found in conventional multivariate data sets but here we focus on their compositional interpretation, as many of these outlier concepts have straightforward geometric interpretation.

6.1.1 Detection of extreme values vs. outliers

A usual way to detect outliers is to compute the Mahalanobis distance of each observation, better if it is based on robust estimates of center and spread. It has been proposed to compare this Mahalanobis distance to the $1-\alpha$ quantile of the appropriate χ^2-distribution, and take as outliers all samples above the corresponding quantile level (Rousseeuw and Leroy, 2003, p 266-270).

An alternative test for the presence of a single outlier can be constructed by using the *maximum Mahalanobis distance* – MDA as test statistics. The distribution of this statistic under the hypothesis of no outlier can be computed by a direct Monte Carlo simulation of normal random samples (Hardin and Rocke , 2005).

The test defined so far is constructed for a single maximum of the robust Mahalanobis distance, if there are several outliers, some sort of correction should be needed. Given that the parameters were estimated in a robust way, the estimates are not influenced by the atypical observations – all observations with a Mahalanobis distance over the $1-\alpha$ quantile of the maximum Mahalanobis distance distribution can be considered as significant outliers. Therefore, no further corrections are required.

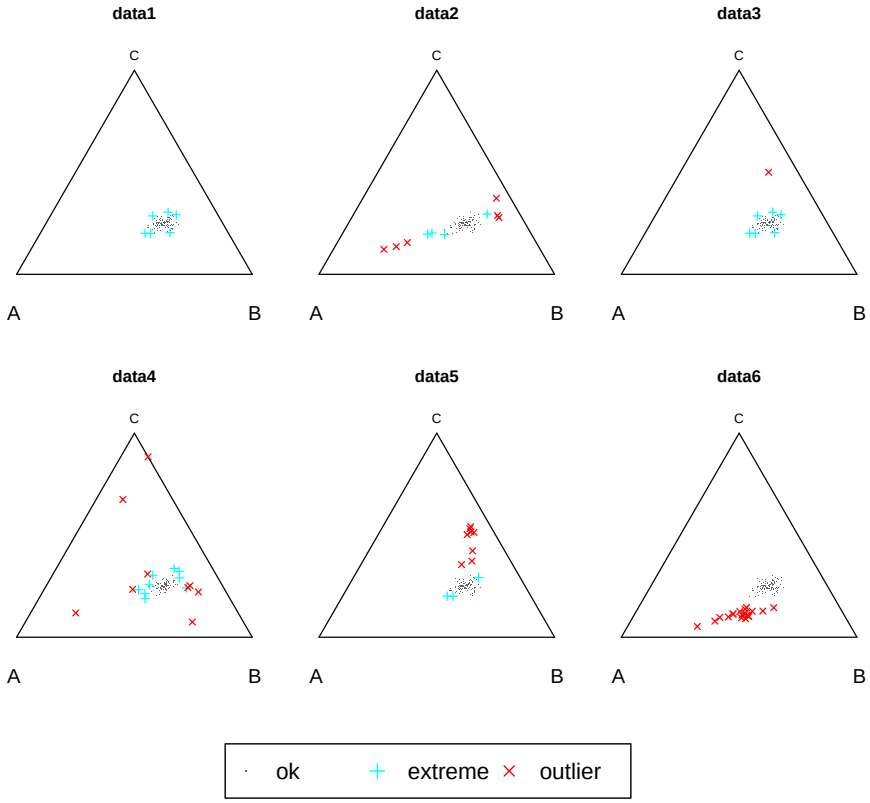

Figure 6: Outlier classification in ternary diagrams of the illustration data sets, using the classification according to the extreme values (type 1) against significant outliers.

Figure 6 shows outliers detected in this way: in red those obtained with the critical value derived from Monte Carlo simulation, and in cyan those detected by comparing with the χ^2 distribution. It can be generated by a commmand like:

```
> outlierplot.acomp(comp, type = "scatter", class.type = "grade")
```

To use the classification based on the χ^2, one must use the optional parameter `corrected=FALSE` in the outlier functions of `compositions`. The implemented approach considers all samples failing any of both tests as *extreme value* (type 1), and those failing the Monte Carlo test as *"true" outliers* which will be further classified through the next sections.

6.1.2 Detection of measurement errors

Outliers generated by a measurement error in a single component should be non-outliers (typical) with respect to the subcomposition of all remaining parts. We could thus formulate D different hypotheses of being a single component outlier in one single components (type 2), and test each one against the alternative of being a general outlier (type 1). All of these tests are built by comparing the Mahalanobis distance for the $(D-1)$-part subcomposition of the candidate outlier with the $1-\alpha$ quantile of the χ^2-distribution with $(D-2)$-degrees of freedom. Obviously, the robust estimates for mean and spread of the studied subcomposition are *plugged* into the Mahalanobis distance.

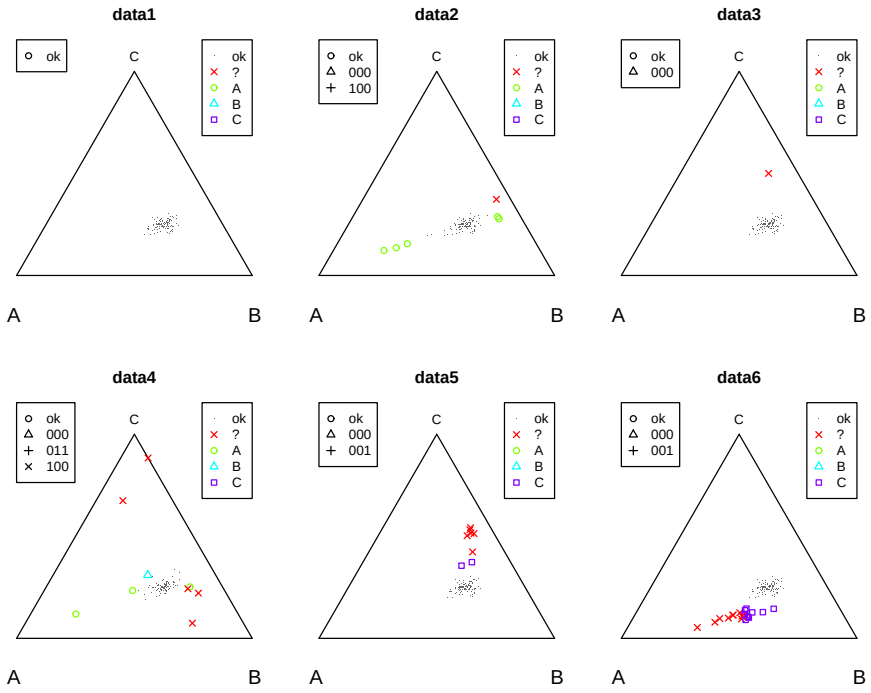

Figure 7: Single component outlier plots of example datasets using color according to the "best" explanation of the outlier, and symbols according to the set of single components explaining the outlier (obtained in a preliminary call with class.type="all") optional argument. This set is displayed by a binary number with one bit for each component from left to right. A 1 says that an error in the corresponding component could explain the outlier: e.g., 101 means that A *or* C offer a sufficient explanation.

In this procedure the hypothesis that an outlier candidate is a single component outlier is rejected, if and only if it is rejected for all D sub-hypothesis. For practical applications, once an outlier has been detected as a single component outlier, one might be interested in the following issues:

- for which parts of the composition was the test passed? In other words, which $(D-1)$-part subcompositions are typical?
- which of these options would give the best explanation? (i.e. the subcomposition without that component is the most typical one, lying the closest in Mahalanobis sense to the robust

average)

- is the wrong component extremely high or extremely low?

A classification according to the first two criteria is shown in Figure 7. It can be generated by a command like:

```
> outlierplot.acomp(comp, type = "scatter", class.type = "best")
```

where the symbols report all possible single-component explanations, and the colour depends on which part provides the best explanation.

6.2 Detection of group outliers

In case of distribution distortion, minor subpopulations and typical contaminations (types 4 to 6) we would probably obtain a number of outliers of the same type with respect to the classifications provided by the previous sections. It would thus be interesting to find out whether these outliers originate from similar processes, come in subpopulations with small spread, or show up independent from each other. Cluster analysis appears as the adequate technique fot this task.

6.2.1 Groups around secondary modes

If we assume that the data set consists of several subpopulations with similar spread (thus, outliers are of type 6), we could try a cluster analysis with complete linkage based on the robustly estimated Mahalanobis distances. To avoid grouping of data with outliers and obtaining too complex dendrograms, it is best to apply the clustering to the outliers only. This method has been proposed and tested by Bren et al. (2007).

6.2.2 Groups on specific directions: contamination and distorted equilibria

The preceding clustering was based on the assumption that subpopulations had small inner Mahalanobis distances, when compared with the global population. On the contrary, contamination outliers are probably around lines, thus having a completely different variance structure. Fortunately, this very hypothesis implies that such outliers might fall in similar directions as seen from the center of the typical population.

Depending on the nature of the process, this direction can be understood either in terms of logratios (e.g. outlier type 2, or outlier type 6 if a chemical equilibrium is locally moved) or in terms of the classical geometry for $\mathbb{R}^d$ (e.g. outlier type 5 when samples are contaminated through mixture, or with a pollutant with more than one part). In both cases we can apply the following idea:

1. choose a geometry: Aitchison for distorted equilibria, classical for (multi-element) contamination,
2. compute the direction from the center of the typical population to all outliers (with respect to the scalar product of the chosen geometry,
3. normalize these differences (using the norm linked to the scalar product),
4. and cluster them with the desired method.

Such a clustering is shown in Figure 8. The directional clustering in Aitchison geometry allows to cluster widely-spread outlier groups according to perturbations in similar directions. The directional clustering in real geometry allows to identify clusters originating from the same contaminant.

The composition of that contaminat must be located somewhere on the line from the center of the population through the center of the contaminated group, but farther than any contaminated sample.

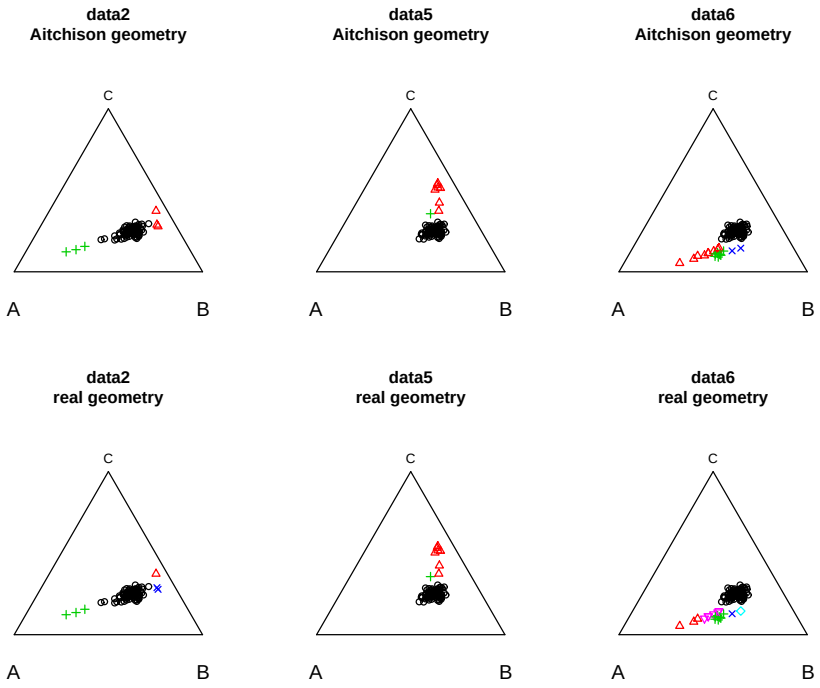

Figure 8: Directional clustering of outliers (complete linkage). Clusters are defined by a cutoff at 15^o. The clustering based on directions in Aitchison geometry nicely identifies the directional outliers in data3 and the secondary group in data6. Clustering in real geometry detects the uniformity of the contamination cluster in data5. No clustering is useful for the small number of outliers in data2 and data3. The heavy tail distribution in data4 showed no structure, and is thus not included. Note that here symbols/colors only display the several groups obtained.

7 Conclusions

Apart from the implementation of the CoDa-dendrogram, we presented two new functionalities dealing with irregular data,

- an elaborated subsystem for dealing with missing data and incomplete observations and
- advance tools for robust statistics, outlier detection and outlier classification

implemented in "`compositions`", the `R` package. Both are based on the global idea that a composition with irregular components is still regular in the remaining subcomposition. Within the scope of a vector space, this idea is translated onto projections: an individual observation with some missing components only specifies the actual position of the datum in the projected subspace corresponding to the observed subcomposition; or an atypical observation in some components is still typical when the whole data set is projected onto that subcomposition subspace. This approach allows for a consistent treatment of outliers and missing values.

In this way, we introduced missing values into graphics, basis definition, and descriptive statistics for compositional data. For presentation purposes, we propose three different ways to plot missing values in order to distinguish their kind and visually assess their relation to the regular part of the data set. In ternary diagrams, all missings are plotted as symbols on the sides joining the

two observed components, or not plotted at all if two parts are absent. Moreover, missing-at-random (MAR) and below-detection-limit (BDL) values are represented as isoproportion lines and semi-lines, respectively, the latter up to the detection limit. Note that, as the simplex is a bounded representation, these lines and semi-lines in fact look like segments. In log-ratio scatterplot representations, MAR and BDL are represented as lines and semi-lines (and here it is apparent that they are not segments), whereas structural zeroes (SZ) pose a difficulty: we propose to plot them as markers on an axis passing through the average of the data set. Finally, when representing data sets with $D \geq 4$ parts, it might be interesting to distinguish those individuals with SZ or BDL on parts that are not being used in the current display.

Regarding the computation of statistics in the presence of missing values, our approach *implies* a different estimation procedure, which is a complement to the replacement technique. This procedure can be summarized in: (i) project each datum to the subspace where it is fully observed; (ii) keep track of the "weight" of each datum on the several possible subspaces (according to the non-missing data); (iii) compute the statistic in each subspace; and (iv) obtain a compatible global statistic from the projected ones by properly down-weighting those with less observations. This method is particularly suited for MAR and SZ types of missing values, but should be used with caution in the case of BDL (in this case, the replacement technique works better). A further technique has been presented, based on selecting a basis grouping together those parts that are typically missed simultaneously: in such bases, coordinates may be computable or replaceable by their mean, and only a few of them behave as censored values.

The robust statistics and outlier detection is based on the minimum covariance determinant approach. The outlier classification is based on several different models of outlier development (measurement errors, equilibrium distortion, contamination of mixture, and the existence of subpopulations), each characterizing an affine subspace related to the *non-atypical* subcomposition. By posing several hypotheses on the distance between the atypical datum and each of these subspaces one can develop Monte-Carlo-based tests to choose the best explanation for each outlier obsrevation. The package provides special plots to represent the resulting classification, which help to understand the nature of outliers in the dataset. In this article we presented ternary diagrams where outliers are classified according to the best explanation, and/or according to all possible explanations. However, the package contains many additional representations, like biplots based on the robust covariance structure, dendrograms showing the robust Mahalanobis similarity between the outliers, distributional representations of the Mahalanobis norm of the data set, etc.

We are aware of the fact that our list of possible causes for atypicality is not complete, but since we provide visualization tools for this classification, the analyst can see if a given sample or group of samples does not satisfy any of these patterns, and than look for possible alternative causes.

Acknowledgements

Funding support to the authors, from the Slovenian Ministry of Science and Higher Education (research program grant P1-0294), and the Department of Universities, Research and Information Society (DURSI, grant 2005 BP-A 10116) of the *Generalitat de Catalunya*, is gratefully acknowledged.

REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96*(456), 1205–1214.

van den Boogaart, K. G., R. Tolosana-Delgado, and M. Bren (2006). Concepts for handling zeroes and missings in compositional data. In Pirard, E., Dassargues, A. and Havenith, H.B. (Eds.), *Proceedings of IAMG'06 — The XI annual conference of the International Association for Mathematical Geology*. University of Liege, Belgium. CD-ROM.

van den Boogaart, K. G. and R. Tolosana-Delgado (2008). "compositions": a unified R package to analyze Compositional Data. *Computers and Geosciences 34*(4), 320–338.

van den Boogaart, K. G., R. Tolosana-Delgado, and M. Bren (in prep). Robustness, classification and visualization of outliers in compositional data. To be submitted to *Computers and Geosciences*.

Bren, M. R. Tolosana-Delgado, and K.G. van den Boogaart (2007). Robustness in compositional data analysis. In: Z. Pengda, F. Agterberg and Q. Cheng (Eds.), *Proceedings of IAMG'07 — The XII annual conference of the International Association for Mathematical Geology*. China University of Geosciences, Beijing, China.

Egozcue, J. J., V. Pawlowsky-Glahn (2005). Coda-dendrogram: A new exploratory tool. In G. Mateu-Figueras and C. Barceló-Vidal (Eds.), *Compositional Data Analysis Workshop – CoDaWork'05, Proceedings*. University of Girona, Spain. CD-ROM and online version: http://ima.udg.es/Activitats/CoDaWork05/

Egozcue, J. J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology 37*(7), 795–828.

Egozcue, J. J. and V. Pawlowsky-Glahn (2006). Simplicial geometry for compositional data. Number 264 in Special Publications, pp. 145–160. The Geological Society, London, UK.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Filzmoser, P. and K. Hron (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences 40*(3), 233–248.

Hardin, J and D.M. Rocke, (2005) The distribution of robust distances, *Journal of Computational and Graphical Statistics*, *14*, 928-946

Huber, P. J. (1981) *Robust Statistics*, Wiley, 308p.

Palarea-Albadalejo, J., J.A. Martín-Fernández and J. Gómez-García (2007). A Parametric Approach for Dealing with Compositional Rounded Zeros. *Mathematical Geology 39*(7), 625–645.

Palarea-Albadalejo, J. and J.A. Martín-Fernández (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences* (in press), doi:10.1016/j.cageo.2007.09.015

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*. Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA) 15*(5), 384–398.

Rousseeuw, P. J. and A. M. Leroy (2003) *Robust Regression and Outlier Detection.*, Wiley, 329p.

Appendix A R instructions used to simulate the data sets

```
# needed packages and accessory functions:
require(rrcov)
require(compositions)
source("FunctionOutliers.R")

# FOR MISSING VALUES:
# desired statistics:
A <- matrix(c(0.4,0.33,0.33,0.3),nrow=2)
MvarMV <- 0.5*ilrvar2clr(A)
McenterMV <- acomp(c(1,2,1))

# the basic non-distorted data set:
dataNMV <- acomp( rnorm.acomp(100,McenterMV,MvarMV) )
colnames(dataNMV)<-c("A","B","C")

# eliminate below the threshold of 5%
dataBDL = dataNMV
 DETLIM = 0.05
 dataBDL[dataBDL<DETLIM]= 0

# eliminate 5% of the data completely at random
dataMAR = dataNMV
 slost = sample(1:length(dataMAR), size=length(dataMAR)*0.05)
 dataMAR[slost] = 0

# FOR OUTLIERS:
# desired statistics
A <- matrix(c(0.1,0.2,0.3,0.1),nrow=2)
Mvar <- 0.3*ilrvar2clr(A%*%t(A))
Mcenter <- acomp(c(1,2,1))

# basic simulation
typicalData <- rnorm.acomp(100,Mcenter,Mvar) # main population
colnames(typicalData)<-c("A","B","C")

# data1: without outliers
data1 <- acomp(typicalData)
 colnames(data1)<-colnames(typicalData)

# data2: with 10% of data with an error in component A
data2 <- acomp( rbind( typicalData +
       rbinom(100,1,p=0.1)*rnorm(100)*acomp(c(4,1,1)) )  )

# data3: with an erratic outlier
data3 <- acomp(rbind(typicalData,acomp(c(0.5,1.5,2))))

# data4: with heavy tails (Cauchy type)
tmp<-set.seed(30)
rcauchy.acomp <- function (n, mean, var){
 D <- gsi.getD(mean)-1
 perturbe(
   ilr.inv(
    matrix(rnorm(n*D)/rep(rnorm(n),D), ncol = D) %*% chol(clrvar2ilr(var))
   ), mean)
}
data4 <- acomp(rcauchy.acomp(100,acomp(c(1,2,1)),Mvar/4))
colnames(data4)<-colnames(typicalData)
```

```
# data5: with an additive pollution through [A,B,C]=c(0.1,1,2)
data5 <- acomp( rbind( unclass(typicalData)+
       outer(rbinom(100,1,p=0.1)*runif(100),c(0.1,1,2)) ) )

# data6: with two subgroups: typicalData,
#     and a newly generated one, around [A,B,C]=c(4,4,1)
data6 <- acomp( rbind(
          typicalData,
          rnorm.acomp(20,acomp(c(4,4,1)),Mvar) ) )
```