Chemometrics and Intelligent Laboratory Systems 143 (2015) 85–96

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Software Description

zCompositions — R package for multivariate imputation of left-censored data under a compositional approach

Javier Palarea-Albaladejo [a,*], Josep Antoni Martín-Fernández [b]

[a] Biomathematics & Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3FD, UK
[b] Dept. Informàtica, Matemàtica Aplicada i Estadística, UdG, Campus Montilivi, Edifici P-IV. E-17071, Girona, Spain

ARTICLE INFO

Article history:
Received 14 October 2014
Received in revised form 20 February 2015
Accepted 24 February 2015
Available online 4 March 2015

Keywords:
Censored data
Nondetects
Zeros
Compositional data
Imputation
Log-ratio analysis

ABSTRACT

`zCompositions` is an R package for the imputation of left-censored data under a compositional approach. It is pertinent when the analyst assumes that the relevant information is contained on the relative variation structure of the data. For instance, in cases where the experimental data are simultaneously measured in amounts related to a same total weight or volume. The approach is used in fields like geochemistry of waters or sedimentary rocks, environmental studies related to air pollution, physicochemical analysis of glass fragments in forensic science, and among many others. In these fields, rounded zeros and nondetects are usually regarded as left-censored data that hamper any subsequent data analysis. The implemented methods consider aspects of relevance for a compositional approach such as scale invariance, subcompositional coherence or preserving the multivariate relative structure of the data. Based on solid statistical frameworks, it comprises the ability to deal with single and varying censoring thresholds, consistent treatment of closed and non-closed data, exploratory tools, multiple imputation, MCMC, robust and non-parametric alternatives, and recent proposals for count data. Key methodological aspects, new contributions, computational implementation and the practical application of the approach are discussed.

1. Introduction

Compositional data analysis refers to statistical theory and methods that focus on the relative variation structure of multivariate data representing relative amounts or portions of a same total. For instance, that total can be the total weight of a solid material or the volume of solutions and gaseous mixtures. They are typically measured in parts per unit, percentages, ppm, moles per litre or the like. The approach assumes that this type of measurement, called compositions in short, carry only relative information. This implies that the total sum of the amounts is regarded irrelevant for the scientific question [1,2]. In practice, most data sets do not contain all possible components and, hence, the observed total is not the actual total amount and it varies between samples. The analyst may decide to re-express the data in e.g. percentages by closing them to add up to 100. This can be useful to, for instance, compare data from samples of different sizes, say sediment samples of 200 and 500 g. But, in any case, it is precisely a main feature of the compositional approach that results do not depend on whether the data are closed to a common constant or not. Although it firstly gained relevance in the field of geochemistry, the use of compositional methods in the diverse disciplines within the natural sciences is rapidly increasing. For example, we can find chemistry-related applications in water research [3], environmental pollution [4], analytical chemistry [5], environmental botany [6], and forensic science [7].

In this work we present a new R package, the `zCompositions`, for the imputation of multivariate data with left-censored values under a compositional approach. The presence of left-censored values complicates any data analysis from the start. Simply discarding them or replacing them by zero may introduce an important estimation bias. We consider what is formally called type I censoring. It is the case where a censoring threshold is given and the number of left-censored values below it is random. Unlike with other types of missing data, we can use that information in the estimation process. Multivariate data analysis techniques such as clustering, principal component analysis, discriminant analysis, and related ones, generally require complete data matrices as input. The imputation of missing or left-censored values by sensible estimates before applying those procedures is an accepted approach [8,9]. In any case, it is worth noting that there is no such thing as an ideal method best performing in all situations. The final choice depends, among others, on the context, the objectives, the plausible assumptions, the nature and size of the data and the number of unobserved values.

A common left-censoring problem in data sets susceptible to a compositional analysis is the presence of rounded zeros [10,11]. That is, small values that have been rounded off to zero due to the number of significant digits considered. When that number is known, it can be

* Corresponding author. Tel.: +44 131 651 7288; fax: +44 131 650 4901.
E-mail addresses: javier.palarea@bioss.ac.uk (J. Palarea-Albaladejo), josepantoni.martin@udg.edu (J.A. Martín-Fernández).

http://dx.doi.org/10.1016/j.chemolab.2015.02.019
0169-7439/

used to define the corresponding censoring threshold. Otherwise, the minimum observed value for the component is often used as such. Moreover, a common practical problem in, for example, applied environmental research is how to handle observations reported to have non-detectable levels of one or several components. Many definitions and interpretations of a detection limit and related concepts have been developed over the years [12]. Following the work by Currie [13, 14] and the United States Environmental Protection Agency (US-EPA), a detection limit (DL) is considered a threshold below which measured values cannot be distinguished from a blank signal, at a specified level of confidence. Note that technological changes or the use of different laboratories, among others, can give rise to multiple DLs for a same component. This is commonly the case in, for example, water chemistry studies. Values near the DL are generally regarded as less precise than those further away. Laboratories often indicate this by providing a limit of quantification (LOQ), which sets a threshold for reliable measurements, usually a multiple of the DL. Nonetheless, values below the LOQ are rarely treated any differently from the rest. In consequence, the analyst has to decide how to appropriately deal with nondetects in data analysis. Simple substitutions by zero, DL/2 or the DL itself have been common practice in environmental or analytical studies [15]. However, concerns have been raised about the impact of such practices on the statistical estimates, particularly when the number of nondetects in a data set cannot be considered negligible.

A number of works, and organisations including the US-EPA, have advocated for methods for censored data exploiting the statistical properties of the data (see for example [16–21]). In regard to software implementations, the `NADA` R package [22] allows obtaining summary statistics for single components including nondetects by using specialised maximum likelihood (ML), robust and nonparametric techniques. Computing regression equations for singly censored data is also possible. Another known freeware alternative is the standalone program `ProUCL`, developed by the US-EPA in Microsoft's .NET Framework for the Windows system only. It offers functionality similar to `NADA` along with additional statistical tools for outlier detection, computing upper statistical limits, statistical tests and useful graphical capabilities. Other general-purpose statistical packages, such as the SAS system, allow for both the statistical analyses and data manipulations required to perform these techniques through macros. The novelty of `zCompositions` is that it considers both the multivariate structure of the data and methods for left-censored data compatible with a compositional approach to data analysis [23].

Using the R statistical programming language [24] implies that the software is open source and multiplatform. The package is freely available for any R system from the CRAN repository (http://cran.r-project.org). The implemented routines build on preliminary code scripts written in MatLab and R by the authors (illustrated in e.g. [25]). Here we introduce a full-featured release within a unified and coherent framework including, among others, new and optimised routines, the ability to deal with varying censoring thresholds, a consistent treatment of closed and non-closed data sets, new visual exploratory tools and some other methodological improvements. Finally, note that `zCompositions` also provide very recent proposals for zeros in count data sets, say e.g. species composition or behavioural data, which are typically treated as multinomial samples. These zeros are assumed to be a consequence of the sampling process, not genuine zeros, and specialised methods are required [26].

In the following, Section 2 describes the strategy of representing the data as coordinates in real space for modelling. Section 3 describes the structure of the package, its distinguishing features, and gives a brief overview of the imputation methods, including new proposals and contributions. Section 4 illustrates the use of the package on data about essential oils composition from *Hyptis suaveolens* plants. Finally, Section 5 concludes with some final remarks and future developments.

2. Some methodological background

The methods in `zCompositions` assume that the observed multivariate data convey only relative information. That is, the focus is on the relative relationships between the components. Thus, the data analysis on a composition $\mathbf{x}$ consisting of D components, $\mathbf{x} = [x_1, \ldots, x_D]$, provides the same answer as one based on a proportional composition $k\mathbf{x} = [kx_1, \ldots, kx_D]$, with k being an arbitrary constant. The relative structure, given by the ratios between components, remains the same regardless of the scale and of whether the observed subset of components is closed or not to a same total sum. This is known as the scale invariance property. Related to this, different analysts can quantify the same set of components in different meaningful units. As long as we can transform the units into each other, for example from mass proportion to volume percentage using the densities of the components, the compositional approach guaranties equivalent results using any possible units. Note that this includes the case of data sets with mixed units.

Another important property, subcompositional coherence, means that results from any subset of components, formally called a subcomposition, and results from the full composition are not contradictory. That is, inferences about some components do not depend on the presence of absence of any other components. Note that this is of great practical relevance as we are almost always working with subcompositions, which the analyst may decide to close or not without affecting the statistical conclusions under a compositional approach. Both representations, closed and non-closed amounts of the same components, are equivalent from a compositional point of view.

The mainstream methodology for a compositional analysis of data is the log-ratio approach. The basic idea is focusing on the ratios x_i/x_j between components, actually on the log-ratios $\ln x_i/x_j$ to enjoy desirable mathematical properties. In the last decade, this has evolved to a more general framework after the characterisation of the simplex, the sample space of compositions, and other subsets of the real space such as the positive real line, as genuine Euclidean vector spaces equipped with their own geometry [27–29]. This allows the data to be isometrically mapped into a space of real coordinates with respect to an orthonormal basis, where standard statistical methods can be used. Results and properties can be transferred back to the original space thanks to the one-to-one transformation linking them. There are infinitely many possibilities to define such an orthonormal basis. Given a particular one, an explicit isometric log-ratio transformation (ilr) can be obtained. For our purposes, the transformation from a composition $\mathbf{x}$ to a vector of coordinates $\mathbf{y}$ with elements given by

$$y_i = \mathrm{ilr}(x_i) = \sqrt{\frac{D-i}{D-i+1}}\ \ln \frac{x_i}{\sqrt[D-i]{\prod_{k=i+1}^{D} x_k}},\quad i = 1, \ldots, D-1, \tag{1}$$

is particularly convenient [11]. For example, suppose a vector of proportions [0.05, 0.09, 0.06], which can be expressed in (closed) percentages as [25,45,30]% on the 3-component simplex. Using Eq. (1), it can be represented on coordinates by the 2-dimensional vector

$$\left[\sqrt{\frac{2}{3}} \ln \frac{0.05}{\sqrt{0.09 \cdot 0.06}},\ \sqrt{\frac{1}{2}} \ln \frac{0.09}{0.06}\right] = \left[\sqrt{\frac{2}{3}} \ln \frac{25}{\sqrt{45 \cdot 30}},\ \sqrt{\frac{1}{2}} \ln \frac{45}{30}\right] = [-0.314,\ 0.287].$$

Fig. 1 displays a simulated data set by means of a ternary diagram (left) and its counterpart in the real space of coordinates according to Eq. (1) (right). The same idea can be applied to univariate data defined on the positive real line $(0, +\infty)$, as it is frequently the case with, for example, environmental observations. In this case, the number e is an orthonormal basis and the real coordinate with respect to it is $\ln x$, which interestingly agrees with the common practice of taking logs of positively valued magnitudes. This characterisation will be used in Section 3.2 to define an imputation procedure based on the lognormal model.

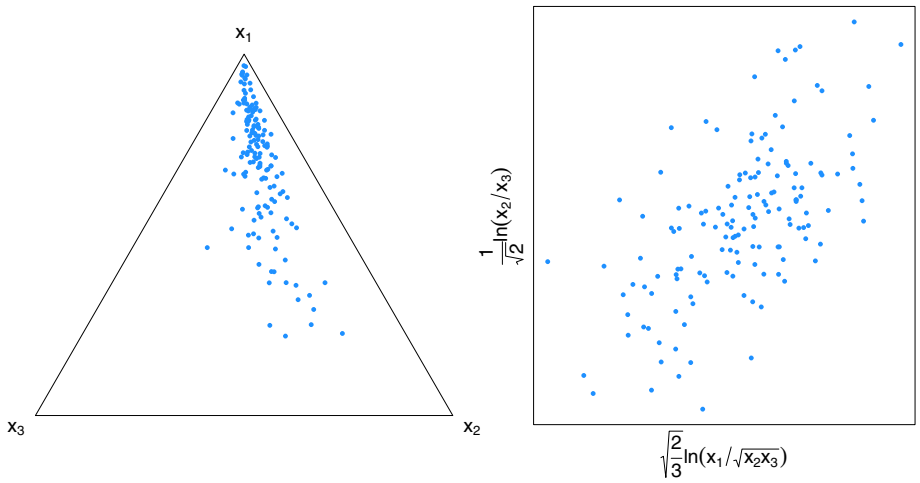

Fig. 1. Simulated compositional data set represented in the 3-component simplex (left) and the real space of ilr-coordinates (right).

We will also employ the so-called additive log-ratio transformation (alr) [1], which relies on log-ratios to a chosen component x_j:

$$\mathbf{y} = \mathrm{alr}(\mathbf{x}) = \left[\ln\frac{x_1}{x_j}, \ldots, \ln\frac{x_{j-1}}{x_j}, \ln\frac{x_{j+1}}{x_j}, \ldots, \ln\frac{x_D}{x_j} \right]. \quad (2)$$

The main drawback of this transformation is that it is not isometric and, hence, the oblique alr-coordinates deform the distances between the original observations. Even so, the alr transformation works properly for statistical modelling and simplifies ordinary estimation in the space of coordinates. Importantly, as shown in [11], analogous results are obtained using either alr or ilr for that purpose.

3. Design and methods

The `zCompositions` package (version 1.0.3 is considered here) was written in R version 3 and benefits from byte-compilation on installation. The main functions currently implemented are summarised in Table 1.

3.1. Workflow

A typical workflow with `zCompositions` is illustrated in Fig. 2. Names in brackets refer to functions, whereas names in parenthesis refer to arguments within functions. All functions take as first argument a rectangular data matrix `X` (class `data.frame` or `matrix`). Then, the unique label used to indicate target values, say nondetects or zeros, must be indicated using the argument `label`.

We can use the function `zPatterns` to obtain a text and graphical representation of the censoring patterns in the data set (example in Fig. 2 using the included example data set `LPdata`) along with some summary statistics. The patterns are displayed into a grid with percentage barplots attached on the margins showing the relative frequencies of left-censored data by component (top) and the relative frequencies of the patterns (right). Note that the patterns are sorted according to their frequency in decreasing order from top to bottom. A vector containing the pattern number for each sample is also generated. A number of graphical parameters allow for customisation of the resulting diagram. From this point a bifurcation is shown in Fig. 2 splitting compositional methods for continuous data and methods for discrete data. The approach to count zeros (`cmultRepl` function, Section 3.2.6) is a recent area of study that diverts from the general left-censoring problem, although the underlying compositional ideas are shared. The argument `dl` is used to enter either a vector or matrix of censoring thresholds. A vector must be provided in the case of single thresholds, one per component. Otherwise, varying thresholds must be entered as a matrix structure of the same dimension as `X`. The cells corresponding to observed values in `X` can actually take any value, filling them in with zeros can be a simple choice. Next, the imputation is carried out according to the chosen method (Section 3.2), and some additional arguments are available. All functions include an initial error checking which controls for correct input specifications.

Table 1
Main functions in the `zCompositions` package.

Name	Short description
`zPatterns`	Summarises and displays the patterns of unobserved values
`lrEM`	Imputation based on the log-ratio Expectation-Maximisation (EM) algorithm
`lrDA`	Single and multiple imputation based on the log-ratio MCMC Data Augmentation (DA) algorithm
`multRepl`	Multiplicative simple replacement
`multLN`	Multiplicative lognormal replacement
`multKM`	Multiplicative KM smoothing spline (KMSS) replacement
`cmultRepl`	Bayesian multiplicative replacement of count zeros

After imputation, the resulting data set (or sets) is scaled to preserve the original format: (a) If the data were originally closed to a constant, then the replaced data set is also closed to the same constant and the components are properly adjusted to preserve the relative structure of the data. (b) If the data were not closed, as it is usually the case in practice, then a re-scaling adjustment as introduced in [25,30] is applied to preserve the relative ratios while the replaced data set is provided in the original scale. The re-scaling procedure works as follows: Let $\hat{x}_j$ be an estimated left-censored value in component j and $\hat{x}_k$, with $k \neq j$, the value given to any other component that was originally observed. Then, the re-scaled estimate is obtained as

$$\hat{x}_j^* = \hat{x}_j \frac{x_k}{\hat{x}_k}, \quad (3)$$

where x_k is the original value of the kth component.

This is included because, whenever data representations given by Eq. (1) or Eq. (2) are applied, the adjustment given by Eq. (3) allows to express the back-transformed data in the same original units. For consistence, we apply the same adjustment even when Eq. (1) or Eq. (2) are not used, for example, in the case of multiplicative replacements (Sections 3.2.3–3.2.5). Note that this might make some censored data to appear as if they had been imputed by values greater than the censoring threshold. However, the preservation of the ratios is guaranteed and it is only a mathematical effect of Eq. (3). Importantly, the resulting data set is compositionally equivalent to the one that would

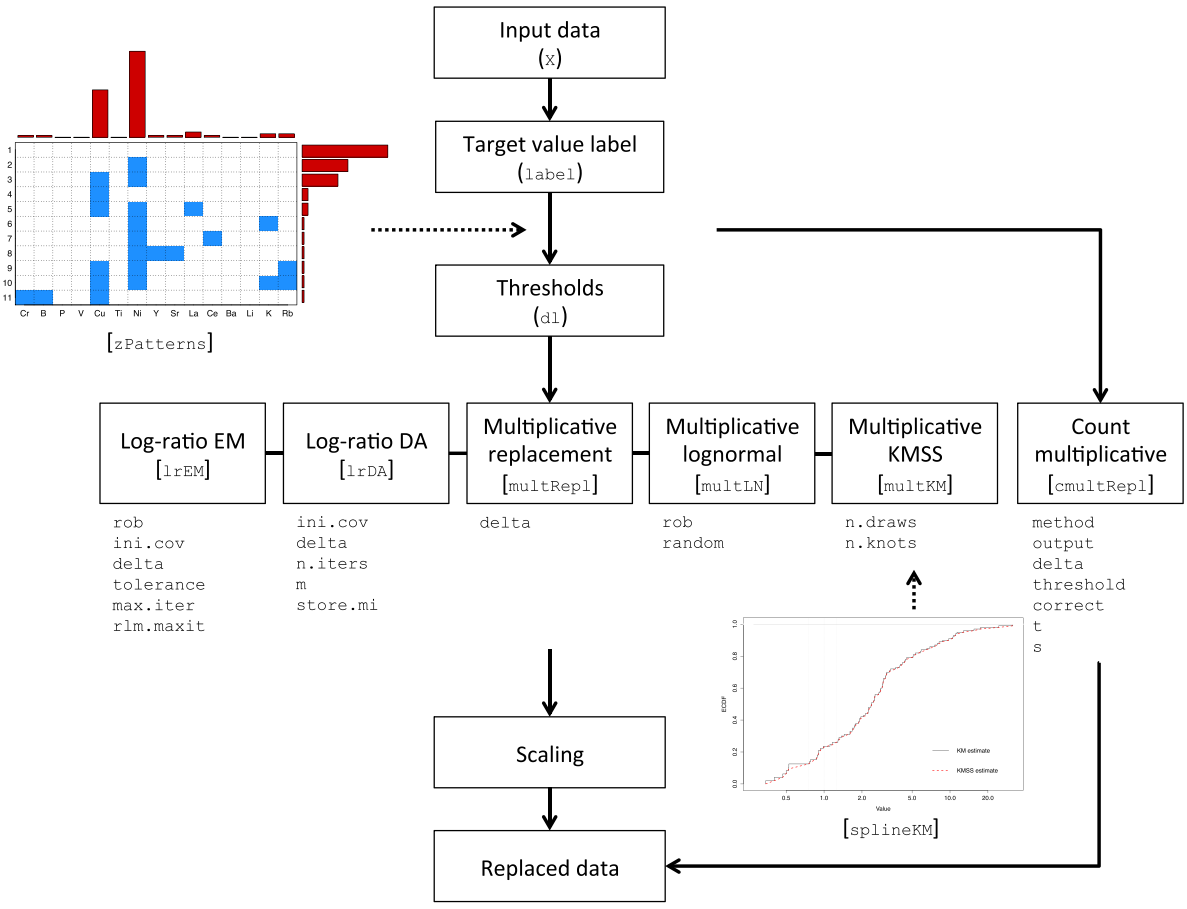

Fig. 2. Structure and typical workflow of `zCompositions`. Names in brackets refer to functions, whereas names in parenthesis refer to arguments within functions.

have been obtained if the data had been initially closed to a constant by the analyst and then imputed, as long as the same scaling had been applied to the censoring thresholds. That is, if the data are for example divided by their sum and multiplied by 100 to turn them into percentages, the same operation must be applied on the censoring thresholds. Note that, when working with a non-closed subcomposition, one could think that simply replacing the censored values by a certain amount, leaving the other components unaltered, would be enough, as no obvious constant-sum must be satisfy and the ratios between the observed components remain unchanged. However this would actually distort the relative structure of the data as illustrated in Section 4.

3.2. Imputation methods

The main core of methods in `zCompositions` are based on the adaptation of current reference statistical iterative imputation frameworks: EM algorithm, Markov Chain Monte Carlo (MCMC) and multiple imputation [8,31]. It is convenient assuming multivariate normality in the space of coordinates, for both technical and scientific practical reasons. Firstly, it facilitates the specification of the conditional relationships between the transformed components while the censoring thresholds are met. Secondly, it is consistent with the characteristic right-skewed distribution of variables representing positive amounts [1,32]. Alternatively, robust and non-parametric approaches are also provided which are tolerant against deviations from that assumption as detailed in the following.

3.2.1. Expectation-Maximisation based algorithm

The Expectation-Maximisation (EM) algorithm [33] is a well-known estimation procedure with missing data. It was adapted to impute rounded zeros in compositional data sets by [34,35]. In a multivariate setting, it uses the information in the covariance structure to produce a conditional estimate of the censored values. Given a censoring pattern with observed, $\mathbf{y}_{obs}$, and unobserved components, $\mathbf{y}_{non}$, the adapted EM iterative scheme comprises basically two steps. At the t*th* iteration:

E-step: Given estimated parameters $\hat{\theta}^{(t)}$, compute $\mathrm{E}\left[\mathbf{y}_{non}|\mathbf{y}_{obs}, \mathbf{y}_{non} < \psi; \hat{\theta}^{(t)}\right]$.
M-step: Find new estimate $\hat{\theta}^{(t+1)}$ based on completed data set $[\hat{\mathbf{y}}_{non}, \mathbf{y}_{obs}]$.

This process is carried out in the space of coordinates, and ψ stands for the mapped censoring thresholds, generically denoted by DL. Assuming multivariate normality of the real coordinates, for each censoring pattern, the conditional expected value of y_{non} is computed at the t*th* iteration in the E step as

$$\hat{y}_{non}{}^{(t)} = \mathbf{y}_{obs}\hat{\beta}^{(t)} - \hat{\sigma}^{(t)}\hat{\lambda}^{(t)}, \quad \text{with} \quad \hat{\lambda}^{(t)} = \frac{\phi\left(\left(\psi - \mathbf{y}_{obs}\hat{\beta}^{(t)}\right)/\hat{\sigma}^{(t)}\right)}{\Phi\left(\left(\psi - \mathbf{y}_{obs}\hat{\beta}^{(t)}\right)/\hat{\sigma}^{(t)}\right)}, \quad (4)$$

where $\hat{\beta}$ and $\hat{\sigma}^2$ are ML estimates of the regression parameters and the conditional variance respectively. The parameter $\hat{\lambda}$, so-called inverse Mills ratio, is the quotient between the standard normal density, ϕ, and distribution, Φ, functions evaluated at the standardised ψ. It accounts for the censoring threshold in order to Eq. (4) to always produce values below it as expected. Under ML estimation, the `lrEM` function (log-ratio EM) actually works in alr-coordinates (Eq. (2)), which provides a fast and numerically stable one-off mapping into the real space. The mapped censoring point is simply $\psi = \ln(DL/x_j)$ in this case. Importantly, the results do not depend on the component x_j used as alr-divisor [35] and, as shown in [11], they are analogous to those

obtained using an isometric mapping via Eq. (1). A distinguishing feature of our implementation is that a correction factor based on the residual covariances obtained using Eq. (4) is applied for the correct estimation of the conditional covariance matrix Σ of the multivariate normal model in the M step. This is required in order to obtain the conditional expectation of the sum of cross-products between two components such that both involve censored values.

Alternately, robust estimation can lessen the potential influence of deviations of the assumptions, mostly due to outlying samples [11]. For this case, working on ilr-coordinates as given by Eq. (1) facilitates robustification and treats outliers consistently. Unfortunately, this comes at the cost of higher computational burden and higher propensity to numerical problems, particularly in complex scenarios as shown in [30]. We have had this in mind and aimed to keep them to a minimum in our implementation. Following [11], adequate one-to-one ilr-coordinates of each censored component are obtained considering a permuted composition $\mathbf{x}^{(j)} = [x_j, x_1, \ldots, x_D]$ such that, sequentially, each censored component x_j is placed in the first position. Then, it turns out that the associated first ilr-coordinate y_1 uniquely represents the censored component x_j with respect to the geometric mean of all the remaining components. An estimated value for y_1 is then obtained by robust MM estimation of Eq. (4) and transferred back to the simplex. Note that, unlike the original proposal in [11], estimated values from preceding EM steps are not used to estimate new ones in `lrEM`.

An initial estimation is required to initiate the EM device. This can be based on either the subset of complete, fully observed samples or a preliminary multiplicative simple replacement (Section 3.2.3). Whereas for ML estimation `lrEM` implements the SWEEP operator [36] as a shortcut to carry out the computations derived from Eq. (4), in the case of robust MM estimation we rely on the `rlm` function (MASS package, [37]). Occasionally, due to both a random element present in the MM method and the feasible range of candidate ilr-values in a particular case, an extremely small ilr-value can be generated which prevents from obtaining a valid covariance matrix. Note that just running the routine once again will do in most cases. Another potential issue is finding samples with only one observed component. If this happens, `lrEM` handles such pattern by multiplicative simple replacement and a warning message is generated.

Finally, note that the `robCompositions` R package [38], centred on robust methods for compositional data, includes alternative implementations of the above procedures for the rounded zero problem.

3.2.2. MCMC data augmentation based algorithm

Data Augmentation (DA) is a Markov Chain Monte Carlo (MCMC) iterative algorithm originally designed for missing data problems [39]. It can be regarded as a Bayesian alternative to maximum likelihood for small data sets, as it bases its inferences on the exact posterior distribution given a particular choice of prior. By the way, the use of priors opens up the possibility of improving the estimations by incorporating external information. Although, for general use, we assume non-informative priors and multivariate normality in the space of coordinates, the procedure can actually come in handy in situations when the likelihood cannot be approximated closely by the normal likelihood. The deterministic E and M steps of the EM algorithm are replaced by simulation-based I (imputation) and P (posterior) steps. At the t*th* iteration:

I-step: Given estimates $\hat{\theta}^{(t)}$, simulate from $\mathrm{P}\left[\mathbf{y}_{non}|\mathbf{y}_{obs}, \mathbf{y}_{non} < \psi; \hat{\theta}^{(t)}\right]$.

P-step: Generate new estimates $\hat{\theta}^{(t+1)}$ by simulating from $\mathrm{P}[\theta|\hat{\mathbf{y}}_{non}, \mathbf{y}_{obs}]$.

Within each censoring pattern, the value $\hat{y}_{non}$ is drawn in the I step from the conditional right-truncated normal distribution with estimated mean $\mathbf{y}_{obs}\hat{\beta}$ and variance $\hat{\sigma}^2$ (being $\hat{\beta}$ and $\hat{\sigma}^2$ as in Section 3.2.1), and truncation point given by ψ. Such a two-stage procedure accounts for the uncertainty in the parameter estimates. The P step involves simulation of the parameters $\theta = (\mu, \Sigma)$ from the completed data posterior given by a normal inverted-Wishart distribution with non-informative priors. Unlike the EM algorithm, the successive estimates of the covariance matrix Σ do not require corrections to the variances. The above I-P sequence generates a Markov chain with the posterior predictive distribution of the (alr) transformed censored data as stationary distribution. Hence, after enough number of iterations, suitable random values can be drawn from the chain to eventually replace them. Algorithms based on the exponential family are generally expected to rapidly reach steady state. This will also depend on the starting point of the process. With this regard, `lrDA` optionally allows for a preliminary run of the `lrEM` routine from which initial parameter estimates near the centre of the posterior distribution are drawn.

The function `lrDA` can also be easily used to implement an MI scheme (see e.g. [40] in the chemometrics context). Basically, by DA-based MI, we obtain m imputed data sets from m independent draws from the generated Markov chain after convergence. In a subsequent data analysis, the variability among the m data sets provides a measure of the uncertainty due to the presence of censored data, which can be combined with ordinary measures of sample variation to lead to single inferential statements, say for example about the proportion of variability explained by the first principal components in principal component analysis. A few draws, say $m \approx 5 - 10$, are often suffice to obtain efficient point estimates [41]. As successive iterates of the Markov chain tend to be correlated, subsampling the chain at regular intervals of length k, large enough for the dependency to be negligible, is suggested. This means that creating m imputations will require km iterations, which will not be necessarily computationally severe in most cases. When MI is chosen, by setting $m > 1$, `lrDA` produces by default a single final data set by averaging the m imputations into efficient point estimates. This may be convenient for subsequent data analysis not including inferential aspects, for example a distance-based clustering analysis. However, the function also allows storing a list of m imputed data sets by setting `store.mi` to `TRUE`. An alternative to MI for propagating imputation uncertainly is using resampling methods such as the bootstrap. A summary of the relative merits of both approaches can be found in, for example, [8]. A compositional bootstrap inference procedure for censored data is depicted in [30].

3.2.3. Multiplicative simple replacement

This method provides a compositional counterpart of the common simple substitution by a fixed fraction of the censoring threshold. It works on the raw data and no coordinate representation is required. The user sets that fraction (`delta` argument) and threshold (`dl` argument), and the censored values in a component x_j are accordingly replaced by `delta*dl`. The distinguishing element is that the remaining components are multiplicatively adjusted to preserve the relative multivariate structure of the data [42]. Such an adjustment for an originally observed component x_k in a closed data set is given by

$$\hat{x}_k = x_k \left(1 - \frac{\sum_{i|x_i<DL_i} \hat{x}_i}{c}\right), \tag{5}$$

where c denotes the given closure constant. When the data are not closed, the resulting data set is adjusted by Eq. (3) to preserve the original scale as described previously. Note that `multRepl` function can also be applied to single compositions.

Based on simulation experiments, [42] recommended using about 65% (`delta` equal to 0.65) of the threshold. We next provide some additional theoretical support to that. Let $f(x)$ be the distribution of a certain component which contains censored values. The left tail of the

distribution is censored in (0, DL), hence the conditional distribution for the observed values is

$$f(x|x<DL)=\frac{f(x)}{\int_0^{DL} f(t)dt,} \qquad 0<x<DL. \tag{6}$$

Consider a random variable y distributed according to a triangular density in $[a, b]$ with mode mod, where $a \le mod \le b$. The expectation of y is $(a + b + mod)/3$. For y defined in the interval [0, DL] and $mod = DL$, the expectation of y is then $2/3DL$, which closely approximates the heuristic proposal $0.65DL$. Assuming that a lognormal with parameters $\mu = 2.5$ and $\sigma = 1$ is a sensible model for the component and $DL = 0.5$, Fig. 3 illustrates such an agreement.

The tails under DL of both, the triangular (dashed line) and the lognormal densities (solid line), are displayed. They intersect by $2/3DL$, that is, 1/3. A Monte Carlo approximation (10^3 random values) of the conditional expectation is $E[x|x < DL] \simeq 0.393$. Additionally, this reflects that the popular $1/2DL$ criterion is not theoretically well supported, as the tail is rarely uniformly distributed.

3.2.4. Multiplicative lognormal replacement

Instead of considering a fixed quantity, this method links the imputed values to the overall characteristics of the data distribution [23]. The lognormal is frequently used to model positively valued data [17,43,44] exhibiting right-skewed distributions. The lognormal mean and variance are traditionally defined through the log-transformed (normal) data mean, μ, and variance, σ^2, as $\exp\{\mu + \frac{1}{2}\sigma^2\}$ and $\mu^2(\exp\{\sigma^2\} - 1)$ respectively [45]. The nonlinearity of these relationships makes lognormal estimation a non-trivial task, and so it has motivated significant research to approximate efficient and unbiased estimators (see e.g. [18,46]). Following the ideas in [28,29], the origin of these difficulties is that the standard lognormal density function is defined with respect to the usual Lebesgue measure in real space. However, this is not indeed a natural measure in the constrained positive real line $(0, +\infty)$. The own Euclidean space structure of this latter implies a different measure from which a normal density can be directly defined on $(0, +\infty)$. While the probability law is the same, the change of representation implies changes in some characteristics of the distribution. Following Section 2, the link between both Euclidean spaces is made by a representation on coordinates, which in this case simply involves working in logarithms and back-transform by exponentiation. This is actually a common practice for obtaining lognormal confidence intervals, which finds proper justification within our approach.

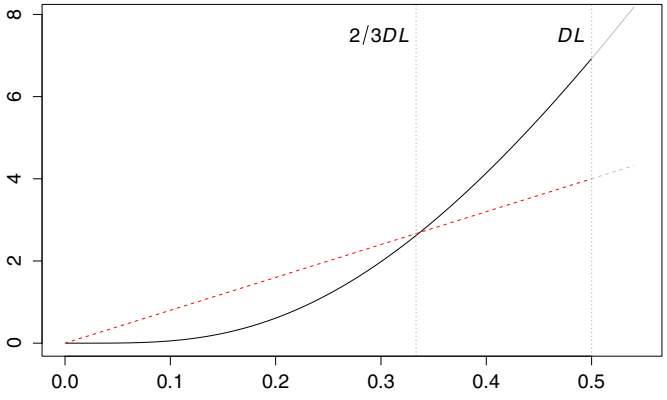

Fig. 3. Lower tail under the limit of detection (DL) of lognormal (solid line) and triangular (dashed line) distributions. Approximate intersection by $2/3DL$.

Given the above, the imputation of censored data implemented in the current version of the `multLN` function is carried out by

$$\exp\left\{\hat{\mu}-\hat{\sigma}\hat{\lambda}\right\}, \quad \text{with} \quad \hat{\lambda}=\frac{\phi((\ln DL-\hat{\mu})/\hat{\sigma})}{\Phi((\ln DL-\hat{\mu})/\hat{\sigma})}. \tag{7}$$

For each component, the estimates $\hat{\mu}$ and $\hat{\sigma}$ are obtained in the log-space of coordinates using methods for censored normal variates. We rely on ML or robust regression on order statistics (ROS) via the `NADA` package. Note however that these estimation methods are used here as part of the process of imputation, and not with the final aim of obtaining estimates of summary statistics of the components. ROS estimation is less dependent on distributional assumptions than ML and, thus, can result most useful for smaller data sets. The expression given by Eq. (7) means that the expected values are estimated by their geometric mean. The remaining components are multiplicatively adjusted by Eq. (5) to preserve the multivariate relative data structure, and the data set is re-scaled by Eq. (3) if it is not closed. Hence, the new approach on coordinates adopted here, apart from being technically sound and coherent, greatly simplifies computations and avoid inconsistencies related to common lognormal inference. Either under ML or ROS, the routine also allows for random imputation from the normal on $(0, +\infty)$ truncated to the interval (0, DL). This feature can also be used for a multiple imputation approach.

3.2.5. Multiplicative Kaplan–Meier smoothing spline replacement

Kaplan–Meier (KM) is a classic non-parametric technique for survival analysis with right-censored data. More recently, it was reformulated for left-censored measurements [21]. It estimates the empirical cumulative distribution function (ECDF), which is then typically used to estimate ordinary univariate statistics. The values for a component (both observed and unobserved) are arranged in decreasing order. The KM ECDF is obtained by computing, at each observed value i, the empirical probability of getting a value less than i. Formally, if n_i denotes the number of values, both observed and censored, at and below i, and d_i is the number of observed values at i, the estimated ECDF is the product of decremental probabilities given by

$$\widehat{\mathrm{ECDF}}(i)=\prod_{i=1}^{k}\frac{n_i-d_i}{n_i}, \tag{8}$$

where k is the total number of observed values. For example, the KM estimate of the mean is computed by integrating the area under the KM curve. KM will assign a probability of 0 to the smallest value. If it happens to be a censored value, the convention is to convert it to the threshold (known as Efron's bias correction). For a single censoring threshold, the KM is equivalent to simple substitution at its value, hence KM is not recommended in that case.

KM imputation has been scarcely found in the literature. In survival analysis, [47] suggest imputing right-censored times, which play the role of our chemical amounts, by randomly drawing from the KM curve among the times of those individuals remaining at risk after the corresponding right-censoring threshold (note that this is in the context for multiple censoring thresholds). Thus, the procedure imputes only observed values beyond that threshold unless the last one is censored, in which case some imputed values may include this last censored value. Turning this into our left-censoring problem, we propose a new way to use the estimated KM ECDF for imputation. We consider a continuous approximation by a cubic smoothing spline. It is used to simulate values below the censoring threshold by inverse transformation sampling from a uniform between 0 and the smoothed $\widehat{\mathrm{ECDF}}$ at DL. The method, KMSS hereinafter, is implemented in the function `multKM`. Unlike [47], it allows to evaluate the approximated ECDF at points other than observed values. Note that, on the basis of the KM

ECDF, the above implies that no values below the minimum in the data are generated. If it is a censored value then KMSS imputes according to the Efron's bias correction. The number of simulated values is controlled by the argument n.points. By default, they are averaged using the geometric mean into a single estimate of each censored value. Alternatively, multKM can produce MI by setting n.points to 1 and executing m runs. Finally, Eq. (5) is applied to preserve the multivariate relative data structure, and the data set is re-scaled by Eq. (3) if it is not closed to be expressed in original units.

Fig. 4 illustrates the procedure by displaying the KM ECDF and its approximation by KMSS for the potassium distribution in the Water data set (available with zCompositions). The three dotted vertical lines represent different censoring thresholds for that component at 0.75, 1 and 1.25 from left to right. This graphical display is produced by the function splineKM and it is useful to determine an optimal smoothing degree for KMSS by setting the number of inner knots the spline function is based on (n.knots argument). It is important to be aware though of the risk of overfitting when too many points are used. Note that multKM allows for different number of knots per component.

3.2.6. Bayesian-multiplicative replacement of count zeros

A compositional count data set is made up of samples comprising discrete vectors of positive values accounting for outcomes falling into any of several mutually exclusive categories. From a statistical point of view, a sample $\mathbf{c} = [c_1, \ldots, c_D]$, with c_j being the number of outcomes corresponding to category j among $n = \sum_k c_k$ trials, is typically considered a realisation of a multinomial distribution. The parameters of this model are $[n; \pi_1, \ldots, \pi_D]$, π_j being the probability of falling into category j. The ML estimates of these probabilities are the proportions $\hat{\pi}_j = c_j/n$. When the relative relationships between the categories are of interest, rather than the total sum n of the vectors, a compositional approach is appropriate. This implies that working with either $\mathbf{c}$ or $\boldsymbol{\pi}$ must be equivalent. In this context, zeros often result from insufficiently large number of trials [48]. The procedure involves Bayesian inference on the zeros and a multiplicative adjustment of the non-zero components.

An imprecise Dirichlet model with parameters s, so-called strength, and $\mathbf{t} = [t_1, \ldots, t_D]$, where $\sum_k t_k = 1$ and the expectation $E[\pi_j] = t_j$, is considered as the prior distribution of $\boldsymbol{\pi}$. The posterior expectation of π_j given that $c_j = 0$, is then

$$E\left[\pi_j | c_j = 0\right] = t_j \frac{s}{n+s} \tag{9}$$

Based on Eq. (9), different settings for s and $\mathbf{t}$ actually define a number of imputation methods provided by the cmultRepl function: geometric Bayesian multiplicative (BM), square root BM and Bayes–Laplace (see [26] for details and comparative analysis). The routine also allows the users to specify their own s and $\mathbf{t}$ parameters. Note that these methods can generate imputed proportions above the lowest estimated probability of a multinomial component. In such cases, the imputation can be corrected by using a fraction (delta) of such minimum. A message informs the user about the number of times this was required. The non-zero components are afterwards multiplicatively adjusted by Eq. (5). Alternatively, multiplicative simple replacement can be directly applied to the matrix of estimated probabilities $\hat{\pi}_j$. For this, an upper threshold, a proportion of $1/n$ as suggested by [49], must be specified. Then, a fraction delta of it is considered to replace zeros. Finally, there is a choice for the output data to be given in proportions, estimated probabilities for each category, or to be re-scaled as compositionally equivalent pseudo-counts.

4. Using zCompositions

We here illustrate the use of zCompositions in practice. The hyptis data set (available in R from the chemometrics package) was originally analysed in [50]. It consists of 30 essential oil samples from *H. suaveolens* (L.) plants at 4 geographical regions in El Salvador: North, South, East-high, and East-low. The concentrations (mass %) of 7 terpenes were analysed by gas chromatography–

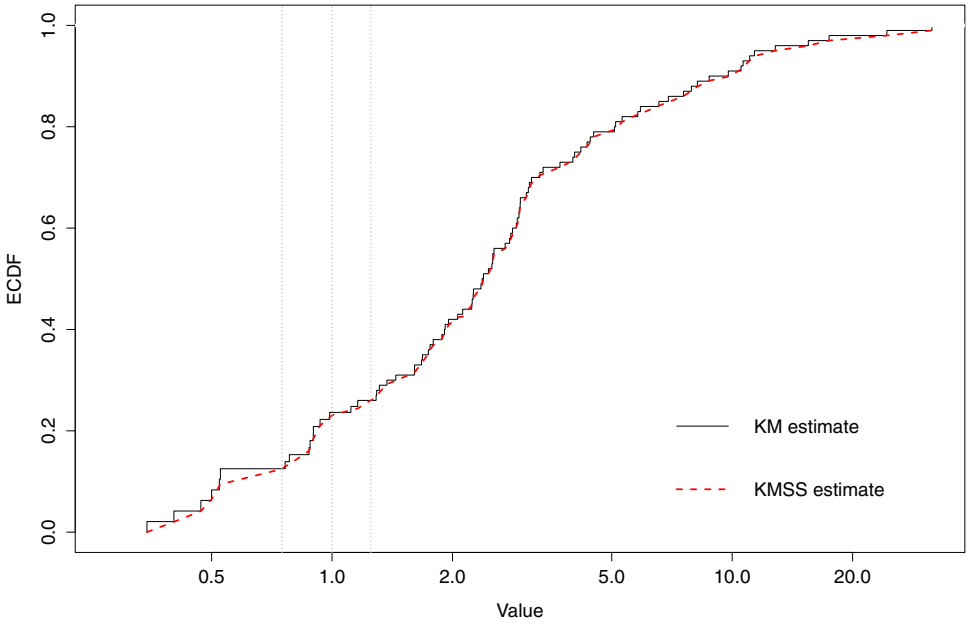

Fig. 4. Empirical cumulative distribution function by KM and KMSS for potassium from the Water data set (x-axis in log-scale). Varying censoring thresholds (vertical dotted lines) at 0.75, 1 and 1.15.

mass spectrometry (GC–MS): Sabiene, β-pinene, 1.8-cineole, γ-terpinene, Fenchone, α-terpinolene, and Fenchol.

```
> data(hyptis,package="chemometrics")
> Hperc <- hyptis[,1:7]
```

The data (stored in `Hperc`) refer to a non-closed data set with zeros assumed to actually indicate non-detected values. The patterns of nondetects can be saved and summarised using `zPatterns`.

```
> Hper.pattern.ID <-
  zPatterns(Hperc,label=0,bar.colors=c("#CAFF70","#A2CD5A"),
          bar.labels=TRUE,cell.colors=c("green4","white"),
          cell.labels=c("Nondetected","Observed"),
          cex.axis=0.8)
```

Fig. 5 shows the graphical output with some customised graphical arguments. The two first analytes, Sabinene and Pinene, are free of nondetects. On the contrary, Terpinolene contains the highest relative amount of them (46.67%). Only 16.67% samples are fully observed, and the most frequent pattern (40% samples) corresponds to those where only Terpinolene contains nondetects.

The analysis reported in [50] includes no details regarding DLs to be used as censoring thresholds. For illustrative purposes, we create an artificial vector of single DLs and consider a 65% fraction (`delta = 0.65`, default) to impute by multiplicative simple replacement.

```
> dlDif <- c(0,0,0.01,0.007,0.004,0.009,0.006)
> Hperc_multReplDif <- multRepl(Hperc,label=0,dl=dlDif)
```

There is only one sample showing pattern #7. The next chunk of code compares the original (sample #26) and imputed compositions.

```
> Hperc[Hper.pattern.ID==7,]
Sabinene Pinene Cineole Terpinene Fenchone Terpinolene Fenchol
9.96 5.5 28.67 0.56 0 0 0

> Hperc_multReplDif[Hper.pattern.ID==7,]
Sabinene Pinene Cineole Terpinene Fenchone Terpinolene Fenchol
9.96 5.5 28.67 0.56 0.002600719 0.005851617 0.003901078
```

Note that the imputed values are not exactly equal to the quantities `delta*dl`. For example, for Fenchone, $0.65 \cdot 0.004 = 0.0026 \neq 0.002600719$. This is due to the adjustment (Eq. (3)) applied to these non-closed data. Importantly, if we close the resulting data to e.g. 100, we obtain the same result as if we had imputed nondetects in a closed version expressed in percentages of the same composition (with DLs accordingly re-scaled). This is illustrated in the following code.

```
> H26repl <- Hperc_multReplDif[Hper.pattern.ID==7,]
> H26repl/sum(H26repl)*100 # Closed after imputation
Sabinene Pinene Cineole Terpinene Fenchone Terpinolene Fenchol
22.28071 12.3036 64.13533 1.25273 0.005817856 0.01309018 0.008726785

> H26 <- Hperc[Hper.pattern.ID==7,]
> H26clos <- H26/sum(H26)*100 # Closed data from the start
> dlDifclos26 <- dlDif/sum(H26)*100 # Same modification to DLs
> multRepl(H26clos,label=0,dl=dlDifclos26) # Imputation
Sabinene Pinene Cineole Terpinene Fenchone Terpinolene Fenchol
22.28071 12.3036 64.13533 1.25273 0.005817856 0.01309018 0.008726785
```

When the composition is originally closed, the replaced values do are exactly `delta*dlDifclos26`. For example, $0.65 \cdot 0.008950548 = 0.005817856$ for Fenchone. As it is desirable, the ratios are preserved in all cases. For example, Sabinene/Terpinolene $= 9.96/0.005851617 = 22.28071/0.01309018$ (except for rounding error). However, this is not the case if we only substitute nondetects in sample #26 by $0.65DL$.

```
> H26sust <- Hperc[Hper.pattern.ID==7,]
> H26sust[5:7] <- 0.65*c(0.004,0.009,0.006)
> H26sust
Sabinene Pinene Cineole Terpinene Fenchone Terpinolene Fenchol
9.96 5.5 28.67 0.56 0.0026 0.00585 0.0039
```

It can be checked that 9.96/0.00585 does not match with the previous results. Next we illustrate ML and robust ROS multiplicative lognormal imputation by `multLN`.

```
> Hperc_multLN <- multLN(Hperc,label=0,dl=dlDif)
> Hperc_multLNrob <- multLN(Hperc,label=0,dl=dlDif,rob=TRUE)
```

It is expected that the more the influence of outliers, the more the differences between imputations. For instance, the estimated geometric means for Terpinolene after both imputations are

```
> exp(mean(log(Hperc_multLN[,6])))
0.1874766
> exp(mean(log(Hperc_multLNrob[,6])))
0.1414918
```

The geometric mean after ROS is lower in this case. By setting `random = TRUE`, we can generate random values from the lower tail of the distribution. This can be used to illustrate differences between both estimation approaches. Focusing on Terpinolene (46% nondetects, $DL = 0.009$) and Fenchol (16.67% nondetects, $DL = 0.006$), we simulated 500 values from their left tails. Fig. 6 shows lower-tail kernel densities estimated from them. It can be observed that the differences are larger for Terpinolene (left) than for Fenchol (right). Interestingly, random ML produces tails looking closer to the typical left tail growing from left to right up to the censoring threshold.

Now we consider iterative estimation using the `lrEM` function. Given that the number of complete samples in `hyptis` (5 samples, pattern #3) is lower than the number of components (7 analytes), a singular covariance matrix is generated as initial estimation. Hence, a preliminary multiplicative simple replacement is carried out instead as suggested by `lrEM`. We show, for example, the estimate of sample #26 for comparison with previous results.

```
> Hperc_lrEM <- lrEM(Hperc,label=0,dl=dlDif,ini.cov="multRepl",
                     max.iter=100)
No. iterations to converge:  55

> Hperc_lrEM[Hper.pattern.ID==7,]
Sabinene Pinene Cineole Terpinene Fenchone Terpinolene Fenchol
9.96 5.5 28.67 0.56 0.0001835043 0.0001541283 0.001495271
```

In this case, the imputed values for sample #26 are notably lower than those provided by the previous methods. Next, we adopt a MCMC approach using the `lrDA` function and taking the estimates from `lrEM` as starting point (this is the default choice, so there is no need to set it up).

```
> Hperc_lrDA <- lrDA(Hperc,label=0,dl=dlDif)
> Hperc_lrDA[Hper.pattern.ID==7,]
Sabinene Pinene Cineole Terpinene Fenchone Terpinolene Fenchol
9.96 5.5 28.67 0.56 0.003452339 0.006820816 0.005447628
```

Alternatively, we can consider MI by setting the argument `m` to e.g. 10. Under the default settings, this implies 10,000 DA iterations picking one estimate out every 1000 steps. Observe that (averaged)

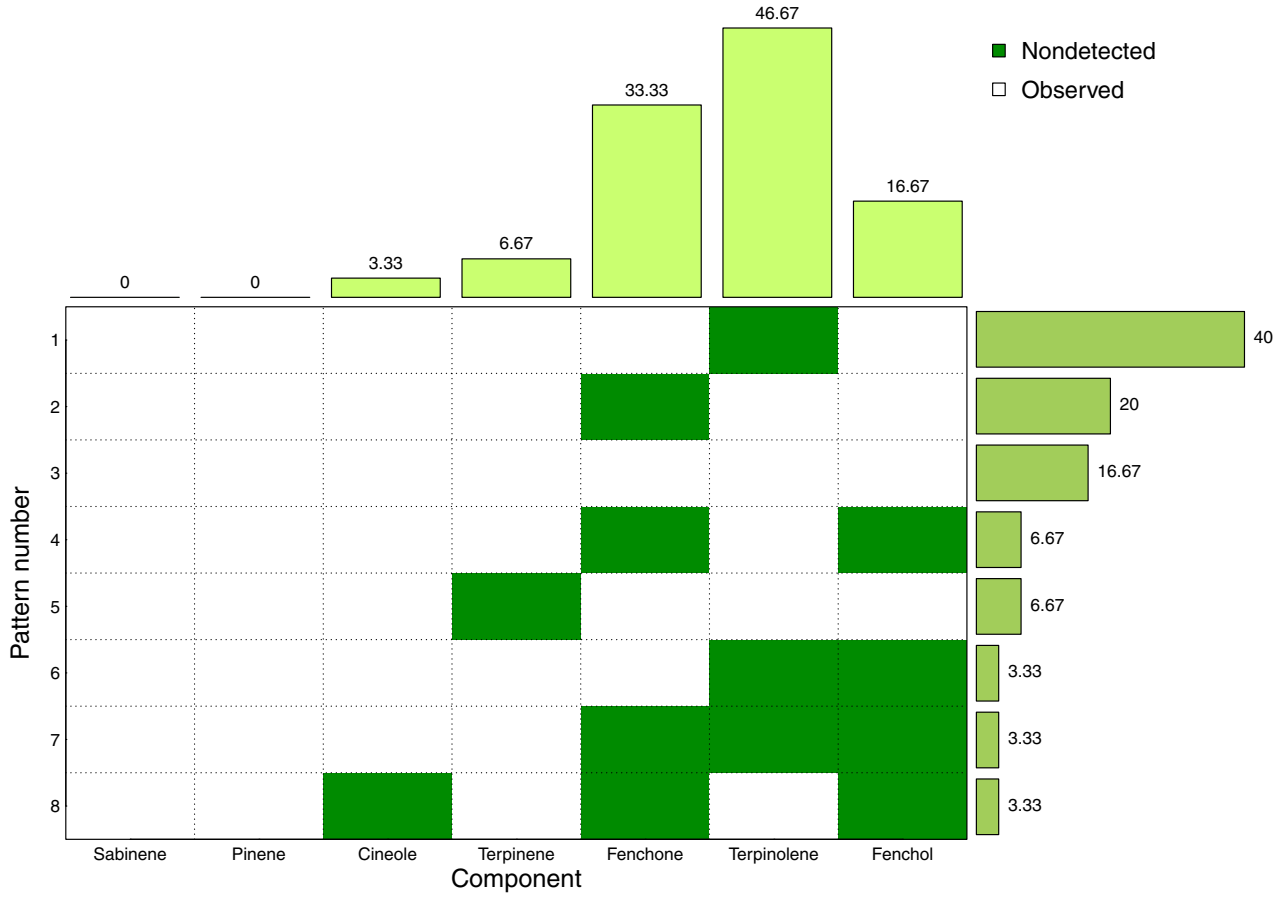

Fig. 5. Patterns of left-censoring in the `hyptis` data set as provided by the `zPatterns` function. Coloured cells in the grid indicate non-detected component within a pattern. Attached barplots display, respectively, percentage number of censored values by component (top) and percentage frequency of patterns (right).

MI provides estimates for sample #26 very similar to single imputation in this case.

```
> Hperc_lrDAMI <- lrDA(Hperc,label=0,dl=dlDif,m=10)
> Hperc_lrDAMI[Hper.pattern.ID==7,]
Sabinene Pinene Cineole Terpinene Fenchone Terpinolene Fenchol
9.96 5.5 28.67 0.56 0.003328096 0.006415999 0.00505668
```

As a summary of the above results, Fig. 7 shows biplots (on log-ratio coordinates) of the imputed data sets.

Observe that the ones from `multRepl` (top-left), `multLN` (top-right) and `lrDA` (bottom-right) are nearly the same. It is `lrEM` (bottom-left) that slightly departs from the general pattern. However, the distribution of the samples and the relative positions of the rays are nearly the same in all cases. The variability explained by these biplots is over 78%, so the data are well represented. The samples from the South region are arranged into a cluster. The samples from the other regions are mixed and show high variability. A contrast between Terpinolene and the pair Fenchone–Fenchol mostly defines the first principal component (horizontal axis). The second (vertical axis) confronts Terpine against the pair Terpinolene–Fenchone.

Finally, note that the commands for the case of varying censoring thresholds would be exactly the same as above but specifying a matrix of censoring thresholds as `dl` argument. Comprehensive documentation and additional examples of all procedures are included in the help pages of the package.

5. Independent testing

This section includes the feedback of the independent testing made on `zCompositions` by Dr David Lovell, lead bioinformatician at CSIRO Australia, and Dr Gary Napier and Dr Tereza Neocleous, researchers at the University of Glasgow.

5.1. Referee 1

`zCompositions` implements a range of imputation strategies for zeros and values below the limit of detection arising in compositional data, along with means to assess the patterns of these zeros. The package is well-suited to the analysis of concentrations as one might find in geochemical or environmental studies. Of critical importance it is the fact that `zCompositions` is founded upon the log-ratio approach pioneered by John Aitchison and, as such, respects the multivariate nature of the data and regards the relative amounts of different components as primarily important.

I downloaded, installed and evaluated `zCompositions` (version 1.0.3) under R (version 3.1.2) with a water chemistry data set that consisted of 141 observations of 18 components out of which 662 (26%) values were zero. Immediately, the `zPatterns` function proved valuable by identifying components and observations where there were very few non-zero values. After estimating the actual detection limits (using the single significant digit of the smallest non-zero value of each component) I ran all of the imputation methods on these data (with the exception of `cmultRepl` which I tested on the Pigs dataset included in the package). All methods executed successfully—`lrEM`

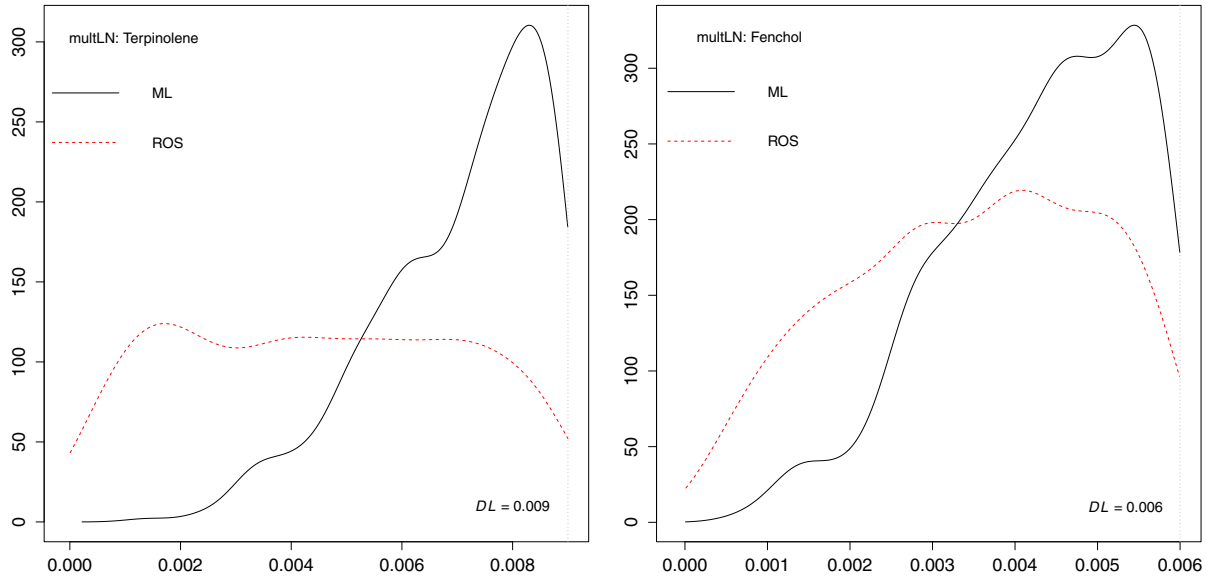

Fig. 6. Kernel density functions estimated from 500 simulated values from the lower tail of components Terpinolene (left) and Fenchol (right) using ML (solid line) and robust ROS parameter estimates (dashed line). The assumed censoring thresholds (dotted vertical lines) are, respectively, 0.009 and 0.006.

and `lrDA` required some observations to be dropped from the data to ensure that at least one component was present in all observations.

The authors have crafted a very useful and straightforward package to enable the principled imputation of zeros in compositional data. Execution was fast with maybe the exception of the iterative `lrEM` and `lrDA` methods in relative large data sets with many different patterns of zeros, which would benefit from some kind of progress indicator and speed-ups in future releases. I recommend `zCompositions` to all analysts who face the challenge of zeros and values below the limit of detection in compositional data.

Independently tested by
Dr David Lovell
David.Lovell@csiro.au
CSIRO, Digital Productivity Flagship
Canberra, Australia

5.2. Referee 2

The `zCompositions` package offers a comprehensive range of options for replacement of nondetects and rounded zeros in compositional data. The available methods can be applied to closed and non-closed compositions, and there are also options to deal with counts (multinomial data). Every method of zero replacement implemented in the package preserves the principles of invariance and subcompositional coherence.

The package is easy to install in the freely available R statistical software, which works on all platforms. We have tested the package using the examples in the very well written help files and also our own compositional datasets. We were impressed with the excellent visual exploratory tools for inspecting zero patterns in compositional data through the `zPatterns` function. The various replacement methods run very fast, even `lrDA` which is implemented using MCMC.

Overall the package is well written and well documented and it provides valuable tools for replacing zeros in compositional datasets in a principled way.

Independently tested by
Dr Gary Napier
g.napier.1@research.gla.ac.uk
and
Dr Tereza Neocleous
tereza.neocleous@glasgow.ac.uk
School of Mathematics and Statistics, University of Glasgow
Glasgow, UK

6. Final remarks and future developments

We have introduced an R package that implements a compositional approach to the imputation of left-censored values in multivariate data sets representing relative portions of a whole. Unlike standard approaches, it considers aspects such as scale invariance, subcompositional coherence, and preservation of the relative variance structure. These properties are desirable under a compositional approach to data analysis. Note, however, that this may not be appropriate when it is assumed that measures of the content of different components present in a sample or material are independent. `zCompositions` offers methods, all formulated under a coherent and unified approach, based on well-established statistical frameworks for incomplete data. Distinctive features of `zCompositions` include consistent treatment of closed and non-closed data, the ability to deal with varying censoring thresholds, graphical exploratory tools, parametric estimation on coordinates, ML, MCMC, robust and non-parametric alternatives, including novel contributions such as multiple imputation and Kaplan–Meier smoothing spline (KMSS) replacement. All this is supplemented with recent proposals for zeros in compositional discrete count data sets.

The flagship methods in `zCompositions` are model-based iterative methods. As it has been stressed, assuming multivariate normality in the space of coordinates is mathematically convenient and sound when the raw data show the characteristic right-skewed profiles of, for example, environmental and geochemical measurements. Even if the model could be somewhat restrictive or unrealistic, it is effectively applied not to the entire data set but only to its censored part. So the eventual effects of model failure are somehow mitigated, especially when the amounts of missing information are not large. In any case, robust and non-parametric alternatives are also provided allowing `zCompositions` to cover a wide range of practical situations. Robust estimation allows for deviations from strict parametric models and can thus be seen as a compromise between parametric and nonparametric approaches. In the general literature about left-censored data, which is mostly focused on obtaining univariate summary statistics, comparative analyses have been performed between ML (typically based on normal or lognormal models), robust (ROS) and nonparametric (Kaplan–Meier) estimation. They have not however shown

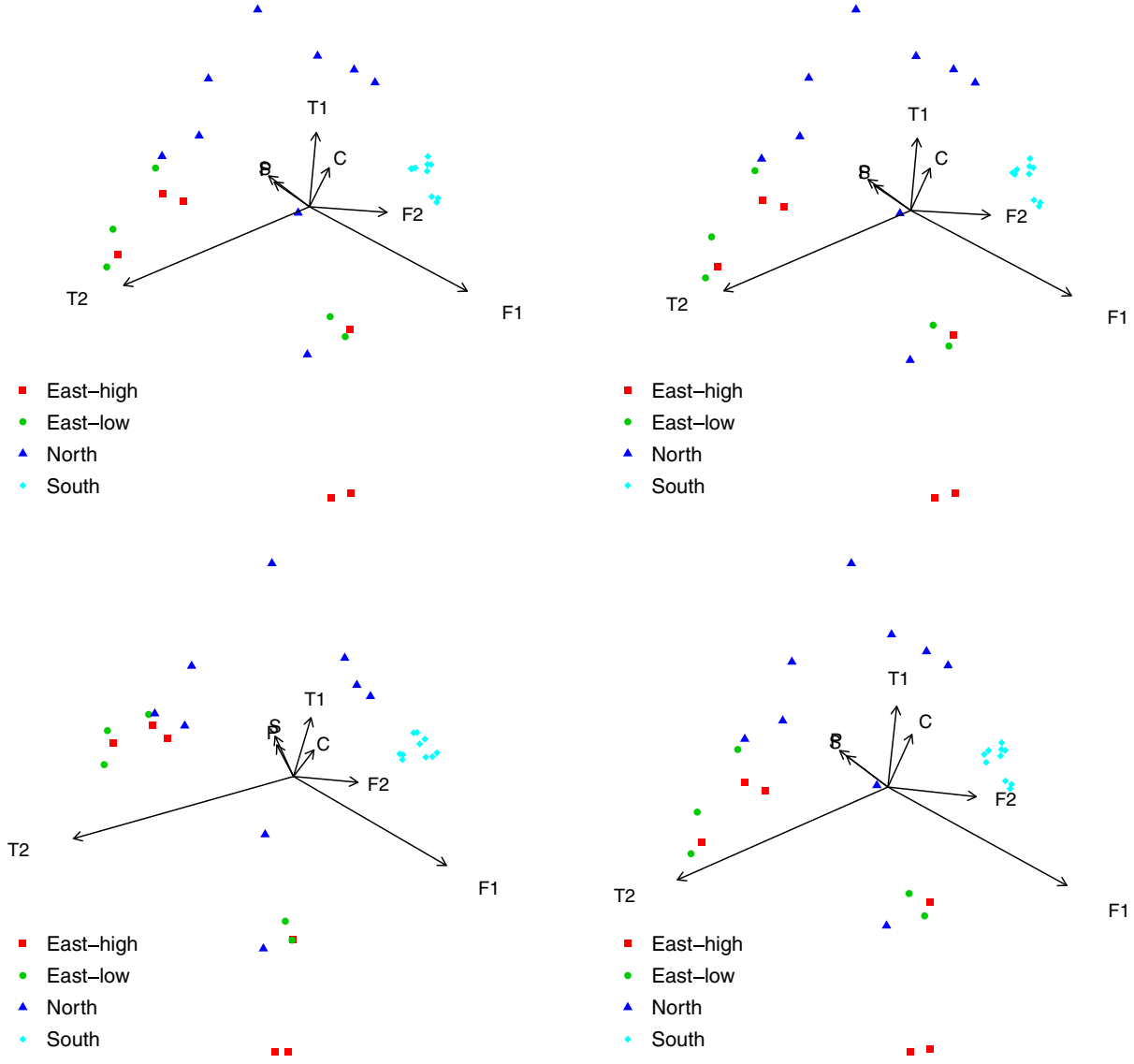

Fig. 7. Log-ratio biplots from the `hyptis` data set after imputation by `multRepl` (top-left), `multLN` (top-right), `lrEM` (bottom-left) and `lrDA` (bottom-right). Abbrev.: S (Sabine), P (Pinene), C (Cineol), T1 (Terpine), F1 (Fenchone), T2 (Terpinolene), and F2 (Fenchol).

one approach consistently outperforming the others. Results depend on testing conditions such as sample size, number of censoring points, departure from the assumed distribution, proportions of censored data, or also the measures of performance used (see e.g. [18–20,51,52]). As to multivariate compositional methods, some comparative performance analyses involving several of the methods included in `zCompositions` have been conducted under different conditions [11,23,30,34,35,53]. Results so far provide overall support to model-based methods. It is important to note with this regard that results from studies on general left-censored data methods are not fully applicable under a compositional approach, as in this case the ability to preserve the relative structure of the data is a key factor which is not generally assessed.

Importantly, the open source nature of the R environment allows extending and adapting the routines according to particular needs or, for example, attaching `zCompositions` as part of another project of more general purpose. It is not new that the blind use of closed software as black boxes with a lot of hidden options may lead to misleading results. Procedures implemented in R instead typically offer full control on parameters and options but, obviously, this also demands a somewhat higher technical background by the user. Like any other well-maintained statistical software libraries, `zCompositions` is planned to evolve by incorporating new functionalities and methodological developments as research progresses in this field and valuable feedback is received from users.

As concluding comments about limitations and related future developments, note that the implemented methods are designed to manage regular data sets where the number of samples (rows) is greater than the number of components (columns). However, new technologies are increasingly generating wide data sets where the numbers of features frequently outnumber by far the amount of samples. New methodological developments are required to cope with this situation in an optimal way. Besides, we have introduced the KMSS as a non-parametric compositional approach to imputation of left-censored data. Further refinements may involve carrying out the fitting process on an adequate space of log-ratio coordinates. Note also that the model-based methods included in `zCompositions` could be easily extended to impute general missing data by lifting the constraint related to the censoring threshold.

Conflict of interest

Finally, the authors state that there is no conflict of interest in relation to this work.

Acknowledgements

This research has been partially supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS), the Spanish Ministry of Economy and Competitiveness under the project "METRICS" (Ref. MTM2012-33236) and the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya (Ref: 2014SGR551). We would also like to thank the three independent testers and anonymous reviewers that kindly contributed with valuable comments to this work.

References

[1] J. Aitchison, The Statistical Analysis of Compositional Data, Chapman and Hall, London, 1986. (Reprinted in 2003 by Blackburn Press).
[2] V. Pawlowsky-Glahn, A. Buccianti, Compositional Data Analysis: Theory and Applications, John Wiley & Sons, Ltd, Chichester, UK, 2011.
[3] N. Otero, R. Tolosana-Delgado, A. Soler, V. Pawlowsky-Glahn, A. Canals, Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a Mediterranean river, Water Res. 39 (2005) 1404–1414.
[4] D. Howel, Multivariate data analysis of pollutant profiles: PCB levels across Europe, Chemosphere 67 (2007) 1300–1307.
[5] M. Korhonová, K. Hron, D. Klimcíková, L. Müller, P. Bednár, P. Barták, Coffee aroma — Statistical analysis of compositional data, Talanta 80 (2009) 710–715.
[6] S. Pignattelli, I. Colzi, A. Buccianti, I. Cattani, G.M. Beone, H. Schat, et al., A multielement analysis of Cu induced changes in the mineral profiles of Cu sensitive and tolerant populations of Silene paradoxa L, Environ. Exp. Bot. 96 (2013) 20–27.
[7] T. Neocleous, C. Aitken, G. Zadora, Transformations for compositional data with zeros with an application to forensic evidence evaluation, Chemom. Intell. Lab. Syst. 109 (2011) 77–85.
[8] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, 2nd ed. Wiley, New York, 2002.
[9] Y. Liu, S.D. Brown, Comparison of five iterative imputation methods for multivariate classification, Chemom. Intell. Lab. Syst. 120 (2013) 106–115.
[10] J.A. Martín-Fernández, J. Palarea-Albaladejo, R. Olea, Dealing with zeros, in: V. Pawlowsky-Glahn, A. Buccianti (Eds.), Compos. Data Anal. Theory Appl, John Wiley & Sons, Ltd, Chichester, UK, 2011, pp. 43–58.
[11] J.A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, J. Palarea-Albaladejo, Model-based replacement of rounded zeros in compositional data: classical and robust approaches, Comput. Stat. Data Anal. 56 (2012) 2688–2704.
[12] D. Coleman, J. Auses, N. Grams, Regulation — From an industry perspective or relationships between detection limits, quantitation limits, and significant digits, Chemom. Intell. Lab. Syst. 37 (1997) 71–80.
[13] L.A. Currie, Limits for qualitative detection and quantitative determination. Application to radiochemistry, Anal. Chem. 40 (1968) 586–593.
[14] L.A. Currie, Detection: international update, and some emerging di-lemmas involving calibration, the blank and multiple detection decisions, Chemom. Intell. Lab. Syst. 37 (1997) 151–181.
[15] I.M. Farnham, A.K. Singh, K.J. Stetzenbach, K.H. Johannesson, Treatment of nondetects in multivariate analysis of groundwater geochemistry data, Chemom. Intell. Lab. Syst. 60 (2002) 265–281.
[16] T. Huybrechts, O. Thas, J. Dewulf, H. Van Langenhove, How to estimate moments and quantiles of environmental data sets with non-detected observations? A case study on volatile organic compounds in marine water samples, J. Chromatogr. A 975 (2002) 123–133.
[17] A. Singh, J. Nocerino, Robust estimation of mean and variance using environmental data sets with below detection limit observations, Chemom. Intell. Lab. Syst. 60 (2002) 69–86.
[18] P. Sinha, M.B. Lambert, V.L. Trumbull, Evaluation of statistical methods for left-censored environmental data with nonuniform detection limits, Environ. Toxicol. Chem. 25 (2006) 2533–2540.
[19] K.F. Leith, W.W. Bowerman, M.R. Wierda, D.A. Best, T.G. Grubb, J.G. Sikarske, A comparison of techniques for assessing central tendency in left-censored data using PCB and p, p'DDE contaminant concentrations from Michigan's Bald Eagle Biosentinel Program, Chemosphere 80 (2010) 7–12.
[20] D.R. Helsel, Much ado about next to nothing: incorporating nondetects in science, Ann. Occup. Hyg. 54 (2010) 257–262.
[21] D.R. Helsel, Statistics for Censored Environmental Data Using Minitab® and R, 2nd ed. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2012.
[22] L. Lee, NADA: Nondetects And Data Analysis for Environmental Data. R Package Version 1.5–6, 2013.
[23] J. Palarea-Albaladejo, J.A. Martín-Fernández, Values below detection limit in compositional chemical data, Anal. Chim. Acta 764 (2013) 32–43.
[24] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
[25] J. Palarea-Albaladejo, J.A. Martín-Fernández, A. Buccianti, Compositional methods for estimating elemental concentrations below the limit of detection in practice using R, J. Geochem. Explor. 141 (2014) 71–77.
[26] J.A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, J. Palarea-Albaladejo, Bayesian-multiplicative treatment of count zeros in compositional data sets, Stat. Model. (2015) (in press).
[27] D. Billheimer, P. Guttorp, W.F. Fagan, Statistical interpretation of species composition, J. Am. Stat. Assoc. 96 (2001) 1205–1214.
[28] J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, C. Barceló-Vidal, Isometric logratio transformations for compositional data analysis, Math. Geol. 35 (2003) 279–300.
[29] G. Mateu-Figueras, The normal distribution in some constrained sample spaces, Stat. Oper. Res. Trans. 37 (2013) 29–56.
[30] J. Palarea-Albaladejo, J.A. Martín-Fernández, R.A. Olea, A bootstrap estimation scheme for chemical compositional data with nondetects, J. Chemom. 28 (2014) 585–599.
[31] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman and Hall, London, 1997.
[32] A. Buccianti, Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach, in: V. Pawlowsky-Glahn, A. Buccianti (Eds.), Compos. Data Anal. Theory Appl, John Wiley & Sons, Ltd, Chichester, UK, 2011, pp. 255–265.
[33] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 39 (1977) 1–38.
[34] J. Palarea-Albaladejo, J.A. Martín-Fernández, J. Gómez-García, A Parametric approach for dealing with compositional rounded zeros, Math. Geol. 39 (2007) 625–645.
[35] J. Palarea-Albaladejo, J.A. Martín-Fernández, A modified EM alr-algorithm for replacing rounded zeros in compositional data sets, Comput. Geosci. 34 (2008) 902–917.
[36] J.H. Goodnight, A tutorial on the SWEEP Operator, Am. Stat. 33 (1979) 149–158.
[37] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, 4th ed. Springer, New York, 2002.
[38] M. Templ, K. Hron, P. Filzmoser, robCompositions: an R-package for robust statistical analysis of compositional data, in: V. Pawlowsky-Glahn, A. Buccianti (Eds.), Compos. Data Anal. Theory Appl, John Wiley & Sons, Ltd, Chichester, UK, 2011, pp. 341–355.
[39] M.A. Tanner, W.H. Wong, The calculation of posterior distributions by data augmentation, J. Am. Stat. Assoc. 82 (1987) 528–540.
[40] M.P. Gómez-Carracedo, J.M. Andrade, P. López-Mahía, S. Muniategui, D. Prada, A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets, Chemom. Intell. Lab. Syst. 134 (2014) 23–33.
[41] J.L. Schafer, Multiple imputation: a primer, Stat. Methods Med. Res. 8 (1999) 3–15.
[42] J.A. Martín-Fernández, C. Barceló-Vidal, V. Pawlowsky-Glahn, Dealing with zeros and missing data in compositional data sets using nonparametric imputation, Math. Geol. 35 (2003) 253–278.
[43] Y. Imaizumi, N. Suzuki, H. Shiraishi, Bootstrap methods for confidence intervals of percentiles from dataset containing nondetected observations using lognormal distribution, J. Chemom. 20 (2006) 68–75.
[44] P.-H. Hsieh, Tales from the tail: robust estimation of moments of environmental data with one-sided detection limits, Comput. Stat. Data Anal. 56 (2012) 4266–4277.
[45] J. Aitchison, J.A.C. Brown, The Lognormal Distribution, Cambridge University Press, London, UK, 1957.
[46] N.T. Longford, Inference with the lognormal distribution, J. Stat. Plan. Infer. 139 (2009) 2329–2340.
[47] J.M.G. Taylor, S. Murray, C.-H. Hsu, Survival estimation and testing via multiple imputation, Stat. Probab. Lett. 58 (2002) 221–232.
[48] P. Walley, Inferences from multinomial data: learning about a bag of marbles, J. R. Stat. Soc. Ser. B 58 (1996) 3–57.
[49] D.A. Elston, A.W. Illius, I.J. Gordon, Assessment of preference among a range of options using Log ratio analysis, Ecology 77 (1996) 2538–2548.
[50] P. Grassi, M.J. Nuñez, K. Varmuza, C. Franz, Chemical polymorphism of essential oils of *Hyptis suaveolens* from El Salvador, Flavour Fragr. J. 20 (2005) 131–135.
[51] P. Hewett, G.H. Ganser, A comparison of several methods for analyzing censored data, Ann. Occup. Hyg. 51 (2007) 611–632.
[52] K. Krishnamoorthy, A. Mallick, T. Mathew, Model-based imputation approach for data analysis in the presence of non-detects, Ann. Occup. Hyg. 53 (2009) 249–263.
[53] A. Buccianti, B. Nisi, J.A. Martín-Fernández, J. Palarea-Albaladejo, Methods to investigate the geochemistry of groundwaters with values for nitrogen compounds below the detection limit, J. Geochem. Explor. 141 (2014) 78–88.