Imputation of missing values for compositional data using classical and robust methods

K. Hron[a], M. Templ[b,c], P. Filzmoser*[,b]

[a] *Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Tomkova 40, 779 00 Olomouc, Czech Republic*

[b] *Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria*

[c] *Statistics Austria, Guglgasse 13, 1110 Vienna, Austria*

Abstract

New imputation algorithms for estimating missing values in compositional data are introduced. A first proposal uses the k-nearest neighbor procedure based on the Aitchison distance, a distance measure especially designed for compositional data. It is important to adjust the estimated missing values to the overall size of the compositional parts of the neighbors. As a second proposal an iterative model-based imputation technique is introduced which initially starts from the result of the proposed k-nearest neighbor procedure. The method is based on iterative regressions, hereby accounting for the whole multivariate data information. The regressions have to be performed in a transformed space, and depending on the data quality classical or robust regression techniques can be employed. The proposed methods are tested on a real and on simulated data sets. The results show that the proposed methods outperform standard imputation methods. In presence of outliers the model-based method with robust regressions is preferable.

Key words: missing values, logratio transformations, balances, robust regression, k-nearest neighbor methods

1. Introduction

Most statistical methods cannot directly be applied to data sets including missing observations. While in the univariate case the observations with missing information could simply be deleted, this can result in a severe loss of information in the multivariate case. Multivariate observations usually form the rows of a data matrix, and deleting an entire row implies that cells carrying available information are lost for the analysis. In both cases (univariate and multivariate), the problem remains that valid inferences can only be made if the missing data are *missing completely at random* (MCAR) (see, e.g., Little and Rubin, 2002). Instead of deleting observations with missing values it is thus better to fill in the missing cells with appropriate values. This is only possible if additional information is available, i.e. only in the multivariate case. Once all missing values have been imputed, the data set can be analyzed using the standard techniques for complete data.

Many different methods for imputation have been developed over the last few decades. While univariate methods replace the missing values by the coordinate-wise mean or median, the more advisable multivariate methods are based on similarities among the objects and/or variables. A typical distance based method is k-nearest neighbor ($k$nn) imputation, where the information of the nearest $k \geq 1$ complete observations is used to estimate the missing values. Another well-known procedure is the EM (expectation maximization) algorithm (Dempster et al., 1977), which uses the relations between observations and variables for estimating

*Corresponding author. Tel.: +43 1 58801 10733, Fax: +43 1 58801 10799

Email addresses: hronk@seznam.cz (K. Hron), templ@statistik.tuwien.ac.at (M. Templ), p.filzmoser@tuwien.ac.at (P. Filzmoser)

the missing cells in a data matrix. Further details, as well as methods based on multiple regression and principal component analysis are described in Little and Rubin (2002) and Schafer (1997). Most of these methods can deal with both, MCAR and *missing at random* (MAR) missing values mechanisms (see, e.g., Little and Rubin, 2002). Moreover, one usually assumes that the data originate from a multivariate normal distribution, which is no longer valid in presence of outliers in the data. In this case the "classical" methods can give very biased estimates for the missing values, and it is more advisable to use robust methods, being less influenced by outlying observations (see, e.g., Beguin and Hulliger, 2008; Serneels and Verdonck, 2008). Classical or robust imputation methods turned out to work well for standard multivariate data, i.e. for data with a direct representation in the Euclidean space (see, e.g., Yucel and Demirtas, 2009). This, however, is not the case for compositional data, and thus a different approach for imputation has to be used.

Compositional data occur frequently in official statistics (tax components in tax data, income components, wage components, expenditures, etc.), in environmental and technical sciences, and in many other fields. An observation $\boldsymbol{x} = (x_1, \ldots, x_D)^t$ is by definition a D-part composition if, and only if, all its components are strictly positive real numbers, and if all the relevant information is contained in the ratios between them (Aitchison, 1986). As a consequence of this formal definition, $(x_1, \ldots, x_D)^t$ and its $c > 0$ multiple $(cx_1, \ldots, cx_D)^t$ contain essentially the same information. A typical example for compositional data are data arising from a chemical analysis of a sample material. The essential information is contained in the relative amounts of the element concentrations, and not in the absolute amounts which would depend on the weight of the sample material. One can thus define the *simplex*, which is the sample space of D-part compositions, as

$$\boldsymbol{x} = (x_1, \ldots, x_D)^t, \quad x_i > 0, \quad i = 1, \ldots, D, \quad \sum_{i=1}^{D} x_i = \kappa. \tag{1}$$

The constant κ represents the sum of the parts. Since only ratios between the parts are of interest, κ can be chosen as 1 or 100, because then the parts of a composition can be interpreted as probabilities or percentages. Note that the constant sum constraint implies that D-part compositions are only $D-1$ dimensional, so they are singular by definition. This causes limitations for the statistical analysis, but the fact that compositional data have no direct representation in the Euclidean space but only in the simplex sample space has even more severe consequences. Applying standard statistical methods like correlation analysis or principal component analysis directly to compositional data would give misleading results (Pearson, 1897; Aitchison, 1986; Filzmoser and Hron, 2009; Filzmoser et al., 2009). This is also true for imputation methods (Bren et al., 2008; Martín-Fernández et al., 2003; Van den Boogaart et al., 2006; Palarea-Albaladejo and Martín-Fernández, 2008).

In this paper we introduce procedures to estimate missing values in compositional data. For this purpose, more details on the nature and geometry of compositional data have to be provided in Section 2. Section 3 introduces an algorithm based on the k-nearest neighbor technique, and an iterative algorithm for the estimation of missings in compositional data. A modification of the latter algorithm will allow to deal with data that are contaminated by outliers. A small data example in Section 4 demonstrates the usefulness of the new routines. In simulation studies in Section 5 the new procedures are compared with standard imputation methods that are directly applied to the raw compositional data. The final Section 6 concludes.

2. Further properties of compositional data

Although compositional data are characterized by the constant sum constraint, the value of κ in Equation (1) can be different for different observations. For example, when sample materials are only analyzed for some chemical elements but not analyzed completely, the sum of the element concentrations of the different samples will in general not be the same. This, however, should not affect the imputation method, because all the relevant information is contained in the ratios between the parts of the observations.

An example is shown in Figure 1 (left). Each data point consists of two compositional parts. The dashed line indicates the constant sum constraint, and according to Equation (1) it is chosen at $\kappa = 1$. The sum of the compositional parts of the data points is smaller than κ. Each point could be shifted along the line from

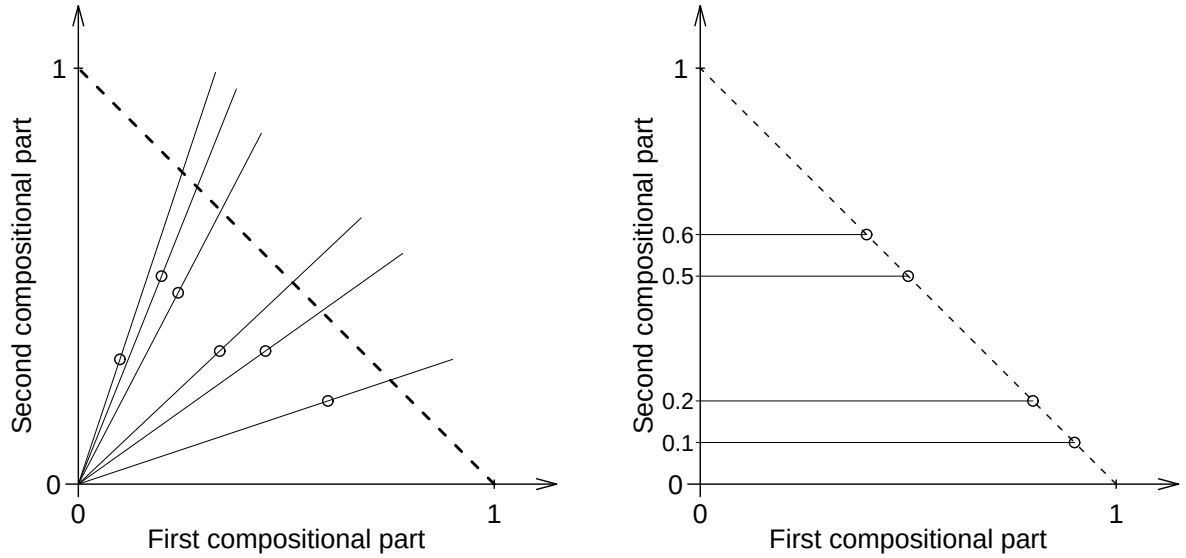

Figure 1: Left plot: Two-part compositional data without the constraint of constant sum. The points could be varied along the lines from the origin without changing the ratio of the compositional parts. Right plot: According to the relative scale, the points close to the boundary are more different than the central points, although the Euclidean distances are the same. The Aitchison distance accounts for this fact.

the origin through the point without changing the ratio of the two compositional parts. More formally, an observed composition $\boldsymbol{x} = (x_1, \ldots, x_D)^t$ is defined as a member of the corresponding *equivalence class* of $\boldsymbol{x}$,

$$\underline{\boldsymbol{x}} = \{c\boldsymbol{x},\, c \in \boldsymbol{R}^+\}.$$

Thus two compositions which are elements of the same equivalence class $\underline{\boldsymbol{x}}$, contain the same information and they are also called compositionally equivalent (Pawlowsky-Glahn et al., 2007). Therefore we could even project the data points along the lines from the origin to the dashed line without changing the information of the compositional data. This fact has to be considered for an appropriate choice of a distance measure.

A further geometrical peculiarity of compositional data is that data points close to the boundary of the sample space are related in a different way than data points in the center. This fact has to be considered for example in outlier detection (Filzmoser and Hron, 2008), but also for designing a distance measure. Figure 1 (right) shows for two-part compositional data two data points close to the boundary and two points in the center of the sample space. Although the data pairs visually have the same distance (because we are thinking in terms of Euclidean distances), the increase along the vertical axis of the boundary points is much larger than that of the points in the center (for the boundary points the increase is by a factor 2 from 0.1 to 0.2, whereas for the central points the increase is only by a factor 1.2 from 0.5 to 0.6). A distance measure that is accounting for this relative scale property is the Aitchison distance (Aitchison et al., 2000), defined for two compositions $\boldsymbol{x} = (x_1, \ldots, x_D)^t$ and $\boldsymbol{y} = (y_1, \ldots, y_D)^t$ as

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

As an example, the two boundary points in Figure 1 (right) have an Aitchison distance of 0.57, whereas the two central points have Aitchison distance 0.29.

Replacing the Euclidean distance by the Aitchison distance is necessary because the simplex sample space has a different geometrical structure than the classical Euclidean space. Principles on this geometry

were introduced in Aitchison (1986), and the resulting so-called Aitchison geometry holds the vector space as well as Hilbert space properties (see, e.g., Egozcue and Pawlowsky-Glahn, 2006). This allows to construct a basis on the simplex, and consequently standard statistical methods designed for the Euclidean space can be used. Out of several proposals for the construction of a basis (Pawlowsky-Glahn et al., 2007), the isometric logratio (ilr) transformation (Egozcue et al., 2003) seems to be the most convenient one. The ilr transformation results in a $D-1$ dimensional real space, and it offers good theoretical and practical properties (Egozcue and Pawlowsky-Glahn, 2005). One important property is the *isometry*, meaning that the Aitchison distance of two compositions $\boldsymbol{x}$ and $\boldsymbol{y}$ is the same as the ordinary Euclidean distance d_E for their ilr images $ilr(\boldsymbol{x})$ and $ilr(\boldsymbol{y})$, i.e.

$$d_A(\boldsymbol{x},\boldsymbol{y}) = d_E(ilr(\boldsymbol{x}), ilr(\boldsymbol{y})). \tag{3}$$

Thus, the ilr transformation allows to represent compositional data in terms of the standard Euclidean geometry, and therefore standard statistical methods can be applied. Note that this property is also fulfilled for the centered logratio (clr) transformation (Aitchison, 1986), but this transformation results in singular data, causing problems for robust estimation. The third well-known logratio transformation, the additive logratio (alr) transformation (Aitchison, 1986) also yields a $(D-1)$-dimensional real space. This transformation, however, is not isometric and thus not recommended for analyzing compositional data (e.g. Pawlowsky-Glahn et al., 2007).

An example is shown in Figure 2. The original two-part compositional data are shown in the left plot with symbols $\circ$, $\triangle$, and $+$. Introducing a constant sum constraint would correspond to projecting the data as indicated, resulting in data with the filled symbols. The ilr transformation reduces the dimensionality by one, and the resulting univariate data are shown on the right-hand side of Figure 2. As it can be seen, the ilr transformation for the original data (upper part) is the same as for the data scaled to constant sum (lower part). The ilr transformed data clearly reveal an outlier group, originating from the data points $+$ in the left plot. Although Figure 2 (left) shows three data clouds, only the cloud with symbols $+$ forms outliers, because the other two clouds cannot be distinguished when thinking in terms of equivalence classes. This fact has to be considered also for the estimation of missing parts in compositional data.

The ilr transformation raises a problem because the new coordinates (often called *balances*) have no interpretation in the sense of the original compositional parts. This is due to the definition of compositional data which contain only relative information, thus making a meaningful interpretation of such variables impossible. A possible solution is to split the parts into separated groups and to construct balances representing the groups and balances representing the relations between the groups. This construction procedure is called *sequential binary partitioning* (Egozcue and Pawlowsky-Glahn, 2005). The resulting balances are in fact just orthogonal rotations, but they have the advantage that groups of variables can be directly assigned to groups of balances.

Sequential binary partitioning is also useful in the context of estimating missing values. For example, if the missing values are mainly contained in the first compositional part of the data, one can choose the ilr transformation as

$$ilr(\boldsymbol{x}) = (z_1,\ldots,z_{D-1})^t, \quad z_j = \sqrt{\frac{D-j}{D-j+1}}\,\ln\frac{\sqrt[D-j]{\prod_{l=j+1}^{D} x_l}}{x_j}, \quad \text{for } j=1,\ldots,D-1\ . \tag{4}$$

This choice of the balances separates all the relative information of the part x_1 from the remaining parts $x_2,\ldots,x_D$. The balance z_1 contains all the relative information of part x_1 to all the remaining parts, represented by $z_2,\ldots,z_{D-1}$. This choice of the balances will be very useful for estimating missing values in x_1 by regression on the remaining variables, see Section 3.

3. Imputation methods for compositional data

In Martín-Fernández et al. (2003) the estimation of missing values in compositional data was done in the sense of the Aitchison geometry, but with the constraint of constant sum of the parts. We follow

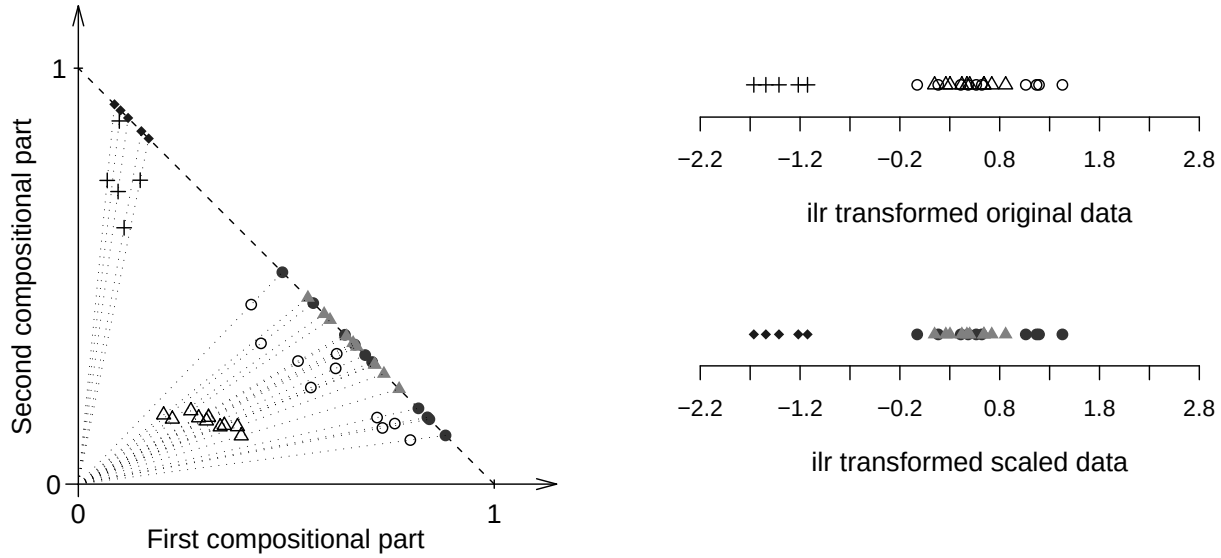

Figure 2: Left plot: Two-part compositional data consisting of three groups, and their projection on the line indicating constant sum 1. While the relative information of the groups with symbols ∘ and △ is similar, the data points of the group with symbol + contain very different information. Right plot: In the upper part the ilr transformed original data are shown. The lower plot shows the ilr transformed data with constant sum constraint. This demonstrates that the constant sum constraint does not change the ilr transformed data.

directly the definition of compositions by considering only ratios between the parts. This is realistic because compositional data include the information in the ratios, and in many practical cases the sum of the parts is not constant (see the example in Section 4). If the constant sum constraint is required, one can divide all data values (the observed and the imputed values) by their sum and multiply by the desired constant. Alternatively, Martín-Fernández et al. (2003) suggested to modify only the nonmissing values.

The easiest way to impute missing values in compositional data is to replace the missing value of a part by the geometric mean of all available data in this part. Here, one has to correct the geometric mean by the ratio of the sum of the parts of the incomplete observation and the sum of the elements of the geometric mean vector (center of compositional data set, see Pawlowsky-Glahn and Egozcue, 2002). This approach, however, does not account for the multivariate data structure, although it is often used for imputation of missing values for compositional data sets. Another approach was introduced by Palarea-Albaladejo and Martín-Fernández (2008). Here the EM-algorithm (Dempster et al., 1977) for alr-transformed compositional data was used for the imputation. Although this algorithm was originally introduced for replacing rounded zeros, it can be adapted for estimating missing values. In the following, this algorithm will be denoted by alr-EM.

In the following we present two new approaches that use the multivariate data information for imputation.

3.1. k-nearest neighbor (knn) imputation

*k*nn imputation turned out to be successful for standard multivariate data (Troyanskaya et al., 2001). The idea is to use a distance measure for finding the k most similar observations to a composition containing missings, and to replace the missings by using the available variable information of the neighbors. In the context of compositional data we have to use an appropriate distance measure, like the Aitchison distance (Section 2).

Suppose that a composition contains missing values in several cells. Then the imputation can be done

(1) simultaneously for all cells, by
   (a) searching the k-nearest neighbors among all complete observations,
   (b) searching the k-nearest neighbors among observations which may be incomplete, but where the information in the variables to be imputed plus some additional information is available;
(2) sequentially (one cell after the other), by
   (a) searching the k-nearest neighbors among observations where all information corresponding to the non-missing cells plus the information in the variable to be imputed is available,
   (b) searching the k-nearest neighbors among observations where in addition to the variable to be imputed any further cells are non-missing.

For estimating the missing parts of a composition, the approaches in (1) use the information of the same k observations, whereas for the approaches in (2) the k observations can change during the sequential imputation. In the following we will use the approach (2a) because in general more neighbors will be considered for imputation, and requesting more information per observation will lead to a more reliable imputation result.

For imputing a missing part of a composition we use the median of the corresponding cells of the k-nearest neighbors. However, we first have to adjust the cells according to the overall size of the parts. This was not necessary for finding the k-nearest neighbors, because the Aitchison distance is the same for any compositions $\boldsymbol{x}$ and $\boldsymbol{y}$ belonging to equivalence classes $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{y}}$ (see Section 2).

More formally, let us consider a composition $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})^t$, $i = 1, \ldots, n$, with n the number of observations, and let $M_i \subset \{1, \ldots, D\}$ denote the set of indexes referring to the missing cells of $\boldsymbol{x}_i$. Then $O_i = \{1, \ldots, D\} \backslash M_i$ refers to the observed parts of $\boldsymbol{x}_i$. For imputing a missing cell x_{ij}, for any $j \in M_i$, we consider among all remaining compositions those which have non-missing parts at positions j and O_i, and compute the k-nearest neighbors $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}$ to the composition $\boldsymbol{x}_i$ using the Aitchison distance. The j-th cell of all k-nearest neighbors is of interest for imputation. First we have to adjust these cells by factors comparing the size of the parts in O_i. The adjustment factors can be taken as

$$f_{ii_l} = \frac{\sum\limits_{o \in O_i} x_{io}}{\sum\limits_{o \in O_i} x_{i_l o}} \qquad \text{for } l = 1, \ldots, k. \tag{5}$$

Using these factors as weights for the observations will make the k-nearest neighbors comparable. The imputed value replacing the missing cell x_{ij} is

$$x_{ij}^* = \text{median}\{f_{ii_1} x_{i_1 j}, \ldots, f_{ii_k} x_{i_k j}\} \; . \tag{6}$$

By taking the median we obtain robustness to outliers in the j-th parts of the k-nearest neighbors.

Although the choice of the adjustments in (5) is coherent with the definition (1) of compositional data, a more robust version could be preferable. In the example and in the simulations below we will thus use the adjustment factors

$$f_{ii_l}^* = \frac{\underset{o \in O_i}{\text{median}}\, x_{io}}{\underset{o \in O_i}{\text{median}}\, x_{i_l o}} \qquad \text{for } l = 1, \ldots, k, \tag{7}$$

which will lead to more stable results for contaminated data.

$k$nn imputation is numerically stable (no iterative scheme is required), but it has some limitations. First of all, the optimal number k of nearest neighbors has to be determined. Ideally, this number can be found within a simulation, by randomly setting observed cells to missing, estimating these missings based on different choices for the number k, and measuring the error between the imputed and the originally observed values. The k producing the smallest error can be considered as optimal. A further limitation concerns small sample sizes. It can happen that when searching for the nearest neighbors using the available information (in fact using subcompositions), the Aitchison distance can lead to nearest neighbors that contain much worse information for estimating the missing values than data points being further away. This, however, could

also happen when using the Euclidean distance for standard multivariate data (Troyanskaya et al., 2001) for searching the nearest neighbors. Therefore, whenenver small sample sizes occur, one has to be aware of this problem with the $k$nn approach. Fortunately, most practical data sets nowadays are of reasonable size.

Finally, $k$nn imputation does not fully account for the multivariate relations between the compositional parts. This is only considered indirectly when searching for the k-nearest neighbors.

From this point of view, the quality of the imputation may be improved by a model-based imputation procedure, as introduced in the following section.

3.2. Iterative model-based imputation

Our focus is on an iterative regression-based procedure. In each step of the iteration, one variable is used as a response variable and the remaining variables serve as the regressors. Thus the multivariate information will be used for imputation in the response variable. Since we deal with compositional data we cannot directly use the original data in regression, but we have to work in a transformed space. For this purpose we choose the ilr transformation with the concept based on balances, because of its advantageous properties. However, already for constructing the balances a data matrix with complete information is needed, see Equation (4). This can be overcome by initializing the missing values with $k$nn imputation, as described above. A further difficulty is that several (or even all) variables have to be used for constructing a balance. Thus, if the initialization of the missings was poor, one can expect a kind of error propagation effect. In order to avoid this, we have to choose the balances carefully. The choice of the balances by Equation (4) is an attempt in achieving the highest possible stability with respect to missing values. For example, the missing values that are replaced in the first variable x_1 will only affect the first balance z_1, but they have no influence on the remaining balances. Thus, using such a sequential binary partition will cause that as few as possible balances are affected by the missing values.

Considering a data matrix with n observations and D parts, Equation (4) can be rewritten for the i-th composition $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})^t$, $i = 1, \ldots, n$, as $ilr(\boldsymbol{x}_i) = \boldsymbol{z}_i = (z_{i1}, \ldots, z_{i(D-1)})^t$, where

$$z_{ij} = \sqrt{\frac{D-j}{D-j+1}}\, \ln \frac{\sqrt[D-j]{\prod_{l=j+1}^{D} x_{il}}}{x_{ij}}, \quad \text{for } j = 1, \ldots, D-1 \quad . \tag{8}$$

The corresponding inverse transformation is $ilr^{-1}(\boldsymbol{z}_i) = \boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})^t$, with

$$x_{i1} = \exp\left(-\sqrt{\frac{D-1}{D}} z_{i1}\right), \tag{9}$$

$$x_{ij} = \exp\left(\sum_{l=1}^{j-1} \frac{1}{\sqrt{(D-l+1)(D-l)}} z_{il} - \sqrt{\frac{D-j}{D-j+1}} z_{ij}\right), \text{ for } j = 2, \ldots, D-1, \tag{10}$$

$$x_{iD} = \exp\left(\sum_{l=1}^{D-1} \frac{1}{\sqrt{(D-l+1)(D-l)}} z_{il}\right) \tag{11}$$

(with possible normalization of the compositional parts to a chosen constant sum).

An iterative algorithm based on regression can be summarized as follows:

Step 1: Initialize the missing values using the $k$nn algorithm based on Aitchison distances, as described above.

Step 2: Sort the variables according to the amount of missing values. In order to avoid complicated notation, we assume that the variables are already sorted, i.e. $\mathcal{M}(\boldsymbol{x}_1) \geq \mathcal{M}(\boldsymbol{x}_2) \geq \ldots \geq \mathcal{M}(\boldsymbol{x}_D)$, where $\mathcal{M}(\boldsymbol{x}_j)$ denotes the number of missing cells in variable $\boldsymbol{x}_j$. Note that $\boldsymbol{x}_j$ denotes now the j-th column of the data matrix.

Step 3: Set $l = 1$.

Step 4: Use the ilr transformation (8) to transform the compositional data set.

Step 5: Denote $m_l \subset \{1,\ldots,n\}$ the indices of the observations that were originally missing in variable $\boldsymbol{x}_l$, and $o_l = \{1,\ldots,n\}\backslash m_l$ the indices corresponding to the observed cells of $\boldsymbol{x}_l$. Furthermore, $\boldsymbol{z}_l^{o_l}$ and $\boldsymbol{z}_l^{m_l}$ denote the l-th balance with the observed and missing parts, respectively, corresponding to the variable $\boldsymbol{x}_l$. Let $\boldsymbol{Z}_{-l}^{o_l}$ and $\boldsymbol{Z}_{-l}^{m_l}$ denote the matrices with the remaining balances corresponding to the observed and missing cells of $\boldsymbol{x}_l$, respectively. Additionally, the first column of $\boldsymbol{Z}_{-l}^{o_l}$ and $\boldsymbol{Z}_{-l}^{m_l}$ consists of ones, taking care of an intercept term in the regression problem

$$\boldsymbol{z}_l^{o_l} = \boldsymbol{Z}_{-l}^{o_l}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{12}$$

with unknown regression coefficients $\boldsymbol{\beta}$ and an error term $\boldsymbol{\varepsilon}$.

Step 6: Estimate the regression coefficients $\boldsymbol{\beta}$ in Equation (12), and use the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ to replace the missing parts $\boldsymbol{z}_l^{m_l}$ by

$$\hat{\boldsymbol{z}}_l^{m_l} = \boldsymbol{Z}_{-l}^{m_l}\hat{\boldsymbol{\beta}}. \tag{13}$$

Step 7: Use the updated balances for back-transformation to the simplex with Equations (9)-(11). As a consequence, the values that were originally missing in the cells m_l in variable $\boldsymbol{x}_l$ are updated. Note that also the non-missing cells are updated, but the ratios between them do not change.

Step 8: Carry out Steps 4–7 in turn for each $l = 2,\ldots,D$.

Step 9: Repeat Steps 3–8 until the Euclidean distance between the empirical covariance matrices computed from the ilr transformed data according to Equation (8) from the present and the previous iteration is smaller than a certain boundary.

Although we have no proof of convergence, experiments with real and artificial data have shown that the algorithm usually converges in a few iterations, and that after the second iteration no significant improvement is obtained.

Note that the choice of the balances by Equation (8) is of advantage in Step 5 in the iterative procedure, because already for $l = 1$, the information of variable $\boldsymbol{x}_1$ with the highest amount of missings is only contained on the left hand side of (12), but not in the explanatory variables on the right hand side.

The estimation of the regression coefficients in Step 6 can be done in the classical way by using least-squares (LS) estimation. However, in presence of outliers we recommend using robust regression which is able to reduce the influence of outlying observations for estimating the regression parameters in Equation (12) (see, e.g., Maronna et al., 2006). In our experiments we used least trimmed squares (LTS) regression because it is highly robust and fast to compute (Rousseeuw and Van Driessen, 2006). Note that robust regression also protects against poorly initialized missing values, because the ilr transformation (8) can lead to contamination of all other cells in the corresponding observation.

4. Numerical study with a data example

The methods described in the previous section are applied to a data set used in Aitchison (1986), p. 395. This data set contains household expenditures on five commodity groups of 20 single men. The variables represent housing (including fuel and light), foodstuff, alcohol and tobacco, other goods (including clothing, footwear and durable goods) and services (including transport and vehicles). Thus they represent the ratios of the men's income spent on the mentioned expenditures. Since this data set is complete, we set the first observation and the third part to missing. The observed value in this cell is 147 HK$ (former Hong Kong dollar).

Additionally, we modify the third observation to see the influence of an outlier on the results of the imputation. Using the procedure of Filzmoser and Hron (2008) one can see that this observation already represents an outlier in the data set. We multiplied the value in the third column by the factors 1 (i.e. the original data set), 2 and 10 to represent a person which is a possible alcoholic. This observation is then

an outlier corresponding to both, the Euclidean and Aitchison geometry. This kind of outlier is denoted as *outlier 1* in the following. In a second outlier scenario, the third row of the original table was multiplied by the factors 1, 2 and 10 to obtain the second kind of outlier which is denoted as *outlier 2* in the following. Depending on the factor, the third person has in general more money to spend, but the proportions remain the same. Therefore this multiplication will not change anything in the Aitchison geometry, but it can affect the estimation in the Euclidean geometry (see Section 2).

The missing value is now estimated with various imputation techniques. The results are shown in Table 1. We apply methods that take the compositional nature of the data into account (geometric mean imputation, iterative LS in the ilr space, iterative LTS (ilr), imputation with the alr-EM algorithm, and $k$nn using the Aitchison distance). For comparison we also use methods that ignore the compositional nature of the data (arithmetic mean, EM algorithm, iterative LS without transformation, iterative LTS without transformation, $k$nn based on the Euclidean distance).

The resulting values in Table 1 demonstrate that methods that account for the compositional nature of the data lead to an improvement of the estimation. Focusing on the *outlier 1* scenario, the more extreme the outlier, the worse the results using standard methods without transformation and non-robust methods in the ilr space (geometric mean and LS (ilr)). For the *outlier 2* situation, the value of the factor does not change anything in the Aitchison geometry. Therefore, the iterative procedures based on LS and LTS regression in the ilr-space, $k$nn imputation based on Aitchison distances, and the alr-EM algorithm give the same results as for the original data. When working in the appropriate space, the model-based procedures are able to improve the initialized values from $k$nn imputation (for $k$nn we used $k = 4$ which gave the best results). The best result is obtained for LS (ilr) with factor 2 in scenario *outlier 1*. Although in this case the outlier has spoiled the regression hyperplane(s), the estimation has been improved by accident, because moving the outlier even further away (factor 10) leads to a severe underestimation.

Table 1: Estimations of the missing value in cell $[1, 3]$ of the expenditures data set (observed value is 147). The considered imputation methods are geometric and arithmetic mean imputation (*gmean* and *mean*), iterative LS and LTS procedure with and without ilr transformation, EM algorithm for alr-transformed (*alr-EM*) and non-transformed (*EM*) data, and $k$nn imputation based on Aitchison and Euclidean distances. *original* corresponds to results for the original data, *outlier 1* is for an outlying observation in both the Aitchison and Euclidean geometries, *outlier 2* is for an outlier only in the Euclidean space. The numbers 1, 2, and 10 are the multiplication factors for generating the outliers. The final values displayed in the table were obtained as proper members of the equivalence class of the corresponding composition using Equation (5).

observed value: 147	*original*	*outlier 1*		*outlier 2*	
method \ factor	1	2	10	2	10
gmean	289.6	300.3	326.9	289.6	289.6
alr-EM	157.8	155.4	150.1	157.8	157.8
$k$nn (Aitch.)	152.1	152.1	152.1	152.1	152.1
LS (ilr)	150.8	148.1	142.2	150.8	150.8
LTS (ilr)	150.8	150.3	150.3	150.8	150.8
mean	330.2	368.4	673.6	368.4	673.6
EM	190.2	214.9	798.4	163.5	195.6
$k$nn (Eucl.)	155.0	155.0	155.0	155.0	155.0
LS (no transf.)	161.0	179.2	324.5	161.3	160.3
LTS (no transf.)	161.3	158.6	158.6	161.4	153.6

Let us remark that by far the worst results are obtained for the univariate imputation methods arithmetic mean imputation and geometric mean imputation, although in the latter case the geometry on the simplex is taken into account. The standard EM algorithm produces more satisfactory results, but the outlier has a big influence on the quality of the estimation.

Although one cannot draw general conclusions from this simple numerical study, it gives a first impression about the performance of different imputation methods. In the next section we will consider more general situations using simulated data. Also a more detailed comparisons between the observed and the imputed

values will be made based on the Aitchison distance and on the covariance structure.

5. Simulation study

For simulating compositional data we will use the so-called normal distribution on the simplex. A random composition $\boldsymbol{x}$ follows this distribution if, and only if, the vector of ilr transformed variables $\boldsymbol{z} = ilr(\boldsymbol{x})$ follows a multivariate normal distribution on $\boldsymbol{R}^{D-1}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (see, e.g., Pawlowsky-Glahn et al., 2007; Mateu-Figueras and Pawlowsky-Glahn, 2008). Thus, $\boldsymbol{x} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes that a D-part composition $\boldsymbol{x}$ is multivariate normally distributed on the simplex. Note that normality on the simplex is independent from the chosen balances for the ilr transformation.

In a first simulation study we generate $n = 100$ realizations from the random variable $\boldsymbol{x} \sim \mathcal{N}_{\mathcal{S}}^{3}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1.05 & 0.95 \\ 0.95 & 1.05 \end{pmatrix}.$$

Using the spectral decomposition, the matrix $\boldsymbol{\Sigma}$ can be re-written as

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (1,1) + 0.05 \begin{pmatrix} 1 \\ -1 \end{pmatrix} (1,-1).$$

Thus, the main variability of the generated data is along one direction, and the variability along the orthogonal direction is small. The resulting covariance matrix thus is poorly conditioned. Although this choice seems quite artificial, situations where only one out of many directions contains an essential portion of variability are realistic in compositional data sets (Kovács et al., 2006). Typical examples are the Arctic lake sediment and Aphyric Skye lavas data sets (Aitchison, 1986, p. 359, 360). Each observation is then multiplied by a factor generated from the uniform distribution $U(0,1)$ on the interval $(0,1)$. This multiplication does not change the equivalence class of the compositional data.

In order to see the influence of outliers on the different imputation methods, $n_1 + n_2$ out of the n observations will be replaced by the following types of outliers:

Outlier group 1: n_1 observations are simulated from $\mathcal{N}_{\mathcal{S}}^{3}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}_1 = (6,0)^t$, and they are multiplied by a factor generated from $U(0,10)$.

Outlier group 2: n_2 observations are taken from the distribution $\mathcal{N}_{\mathcal{S}}^{3}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and they are multiplied by a factor coming from $U(0,10)$.

Within the simulation scheme the numbers of outliers in both groups will be the same, and they will be increased from 0 to 40, i.e. $n_1 = n_2 = 0, 1, \ldots, 40$.

The role of the outlier groups is similar to the two types of outliers in the numerical study of Section 4. *Outlier group 1* consists of observations that are potential outliers in the Euclidean space as well as in the Aitchison space, whereas *outlier group 2* will not have any effect in the Aitchison geometry due to the properties of equivalence classes for compositions (see Section 2).

The amount of missing values is fixed but different for each variable. 20% of the values in the first variable, and 10% in the second variable are set to be missing completely at random. Missing values are only generated in the non-outlying data group (also in the subsequent simulations), because here a fair comparison of different (classical and robust) methods is possible.

The advantage of this design in low dimension is that it is possible to visualize the generated three-part compositions in a ternary diagram. A ternary diagram is an equilateral triangle $X_1X_2X_3$ such that a composition $\boldsymbol{x} = (x_1, x_2, x_3)^t$ is plotted at a distance x_1 from the opposite side of vertex X_1, at a distance x_2 from the opposite side of vertex X_2, and at a distance x_3 from the opposite side of vertex X_3 (see, e.g., Aitchison, 1986). Figure 3 (left) shows a simulated data set where each of the outlier groups represents 5% of the observations. The 5 observations from *outlier group 1* are shown with symbol +, the 5 points from *outlier group 2* have symbol $\triangle$, and the remaining regular points are plotted with small dots. The points + from *outlier group 1* are close to the boundary of the ternary diagram, and the points $\triangle$ from

outlier group 2 are spread over the main data cloud. Note that the points △ do not appear as outliers in the ternary diagram because the different sums of the parts cannot be visualized. Figure 3 (right) shows the two dimensions of the ilr transformed data. Here the shifted center of the points + from *outlier group 1* is clearly visible, whereas the points △ are not acting as outliers in this space.

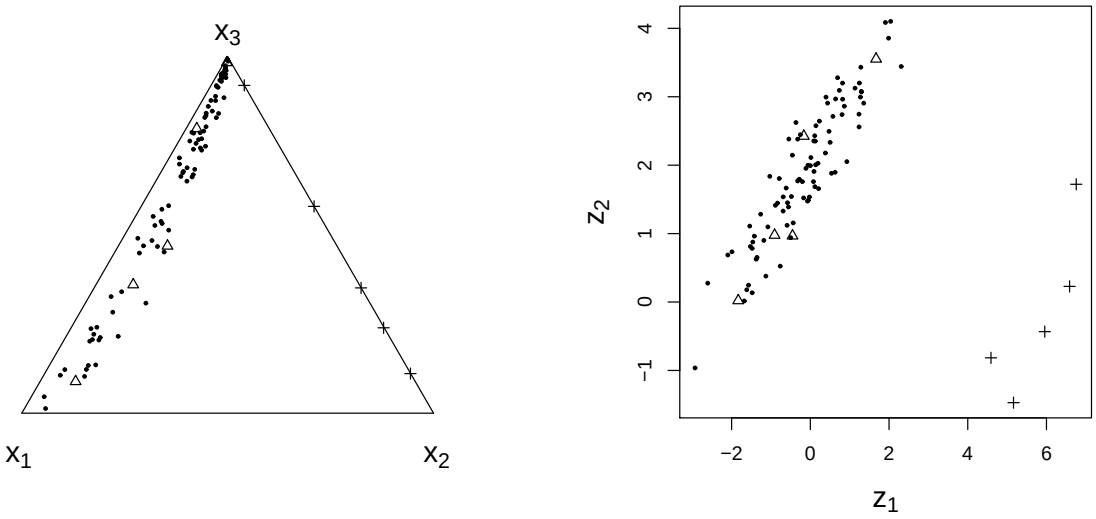

Figure 3: Simulated data set with 5 points from *outlier group 1* (symbol +) and 5 points from *outlier group 2* (symbol △). Left plot: 3-part compositions plotted in the ternary diagram; right plot: data after ilr transformation.

The choice of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the simulation study determines the shape of the raw (compositional) data on the simplex. The further away $\boldsymbol{\mu}$ is from the null vector (indicates equilibrium on the simplex), the closer the compositions are to the border of the simplex. The choice of the covariance matrix $\boldsymbol{\Sigma}$ determines how elongated the data points appear in the ternary diagram.

The original and the imputed data values are compared by two different criteria:

Compositional error variance: Let $M \subset \{1, \ldots, n\}$ denote the index set referring to observations including at least one missing cell, and $n_M = |M|$ be the number of such observations. The *compositional error variance* is defined as

$$\frac{1}{n_M} \sum_{i \in M} d_A^2(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_i) \tag{14}$$

where $\boldsymbol{x}_i$ denotes the original composition (before setting cells to missing), and $\hat{\boldsymbol{x}}_i$ denotes the composition where only the missing cells are imputed (see also Martín-Fernández et al., 2003).

Difference in covariance structure: We denote $\boldsymbol{S} = [s_{ij}]$ as the sample covariance matrix of the non-outlying original ilr transformed observations z_{ij}, using for example the transformation from Equation (8) (another ilr basis would not change our criterion introduced in the following). Further, $\tilde{\boldsymbol{S}} = [\tilde{s}_{ij}]$ denotes the sample covariance matrix computed with the same ilr transformed observations where all missing cells have been imputed. The *difference in covariance structure* is based on the Euclidean distance between both covariance estimations, namely as

$$\frac{1}{D-1} \sqrt{\sum_{i=1}^{D-1} \sum_{j=1}^{D-1} (s_{ij} - \tilde{s}_{ij})^2} = \frac{1}{D-1} \|\boldsymbol{S} - \tilde{\boldsymbol{S}}\|. \tag{15}$$

Thus the *compositional error variance* measures closeness of the imputed values in the Aitchison geometry, whereas the influence of the imputation to the multivariate data structure is expressed by the *difference in covariance structure*.

For each considered percentage of outliers (0% to 40% per outlier group, in steps of 1%) we simulated 1000 data sets according to the above scheme, and estimated the missing values with different techniques. Then the average for the above quality criteria is computed. The results are displayed in Figure 4, and they confirm the findings of the numerical study in Section 4. The left column of the pictures in Figure 4 shows the results for the non-transformed data (the imputation is done in the Euclidean space), the right column is for the transformed data (the imputation is done in the Aitchison geometry). The top row of the pictures shows the average of the *compositional error variance*, the bottom row presents the average *difference in covariance structure*. Since we use the same scale, it is easy to compare the pictures of one row. If the compositional nature of the data is ignored (left column), the results generally get worse. An exception are the $k$nn results for the *difference in covariance structure*, which are similar when using the Euclidean or the Aitchison distance. For kNN the best results were achieved for $k = 8$, but other choices have a rather small effect on the outcome.

When working in the proper geometry (right column of the pictures in Figure 4), the initial kNN imputation can be improved considerably with the iterative model-based imputation. If no outliers are present, the use of LS-regressions or LTS-regressions within the model-based procedure leads to comparable results. However, in presence of outliers the robust imputation algorithm shows clear advantages over LS. The iterative LTS procedure remains very stable up to about 35% outliers.

Both quality measures, the compositional error variance and the difference in covariance structure, are adequate in the context of compositional data. Using corresponding measures based on the Euclidean geometry is inappropriate in this case, which was a reason why we outlined the difference between the Euclidean and the Aitchison geometry in detail. The methods applied in the left column of Figure 4 are based on the Euclidean distance, but still the data are compositional data. Thus, for reasons of comparability with methods for the Aitchison geometry, we kept the same vertical axes as for the pictures in the right column of Figure 4.

Since the previous simulation study is limited to low dimensionality and to a special choice of the covariance matrix, we want to provide deeper insight with more general situations. Thus, in a second and third simulation setup $n = 100$ realizations from the random variable $\boldsymbol{x} \sim \mathcal{N}_S^{10}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are generated (higher dimension). $\boldsymbol{\mu}$ is chosen as the null vector (equilibrium on the simplex), and all diagonal elements of $\boldsymbol{\Sigma}$ are equal to 1. The off-diagonal elements of $\boldsymbol{\Sigma}$ are chosen as 0.9 for the second setup, and 0.1 for the third setup. We thus can expect that the second setup will be advantageous for the model-based procedures, but that the third setup is particularly difficult because of the low correlations. Only outliers of type 1 are considered, and they are sampled from $\mathcal{N}_S^{10}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}_1 = (6, 0, \ldots, 0)^t$. The missing values are generated with MCAR. For the second setup, the percentages of missing values in variables 1 to 10 were chosen as 20%, 10%, 5%, 2%, ..., 2%, 0%, respectively. In the third setup they are taken as 5%, only the last variable is without missings (which is important for the alr-EM algorithm).

Figure 5 shows the results from 1000 simulated data sets corresponding to the second (left column) and the third (right column) simulation configuration. The results of only those imputation methods are shown which take the compositional nature of the data into account; results of the other methods are generally worse, although no outlier group 2 is considered. For the second setup (Figure 5, (a) and (c)), the iterative LS method yields the best results in the uncontaminated case. In presence of up to 30% outliers, the iterative LTS method shows the best performance. Outliers have a large effect on all other methods to both, the compositional error variance and the difference in covariance structure. The most difficult situation, the third simulation setup, leads to a similar behavior of all considered imputation methods (Figure 5, right column). In presence of outliers the iterative LTS method performs slightly better.

Finally, in our simulation the missings were generated according to an MCAR situation. We also tested the MAR situation, leading to analogous conclusions. The reason is that multivariate imputation methods such as kNN and the proposed model-based approach can deal with MAR, but univariate mean imputation techniques which have shown very poor performance already for MCAR cannot deal with the MAR situation

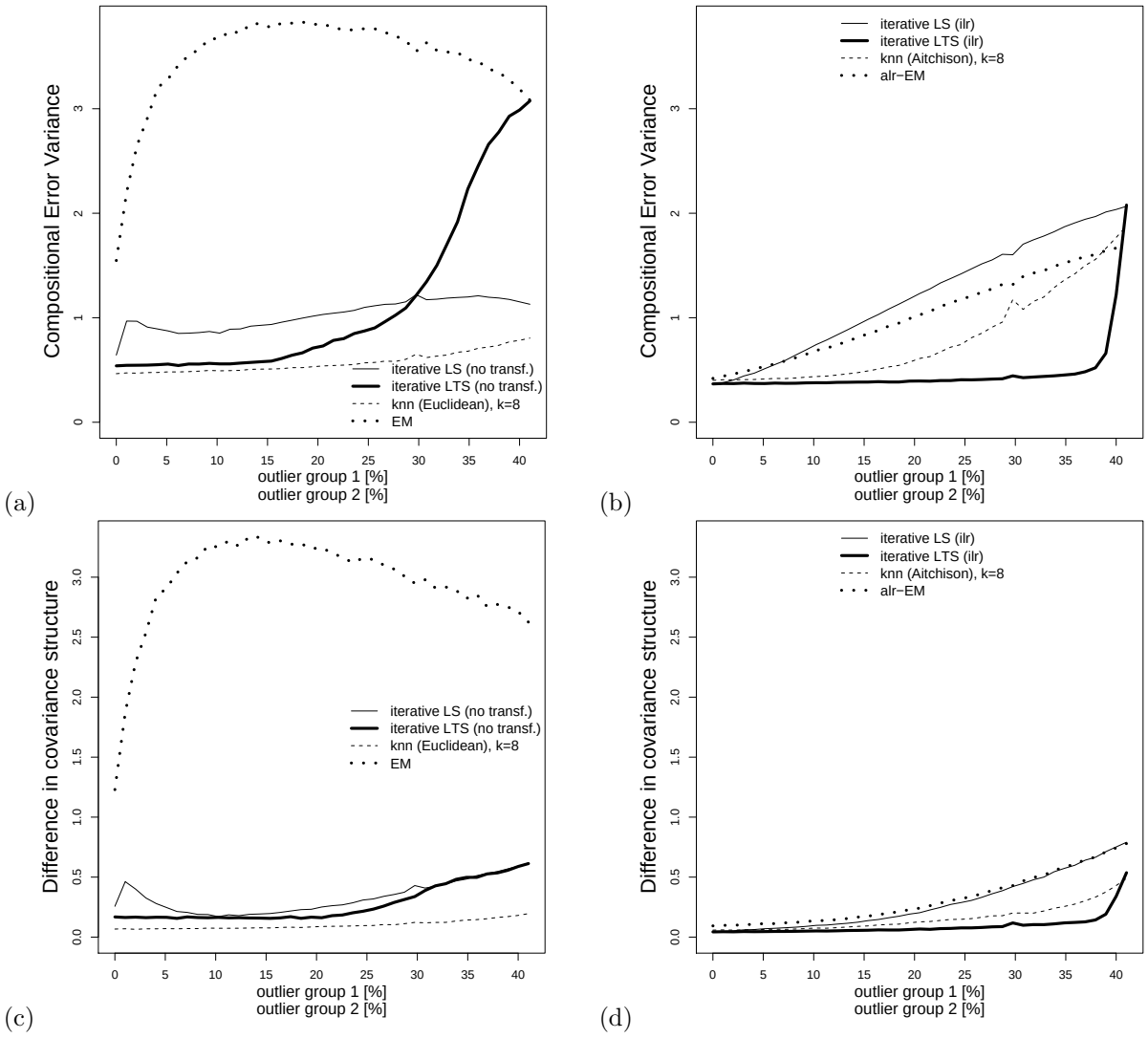

Figure 4: Simulation results (average of the *compositional error variance* and the *difference in variation*) for kNN imputation, model-based imputation using iterative LS and LTS regressions and EM/alr-EM algorithm. For (a) and (c) the imputation is done in the Euclidean space (no transformation), for (b) and (d) the imputation methods are applied in the Aitchison geometry.

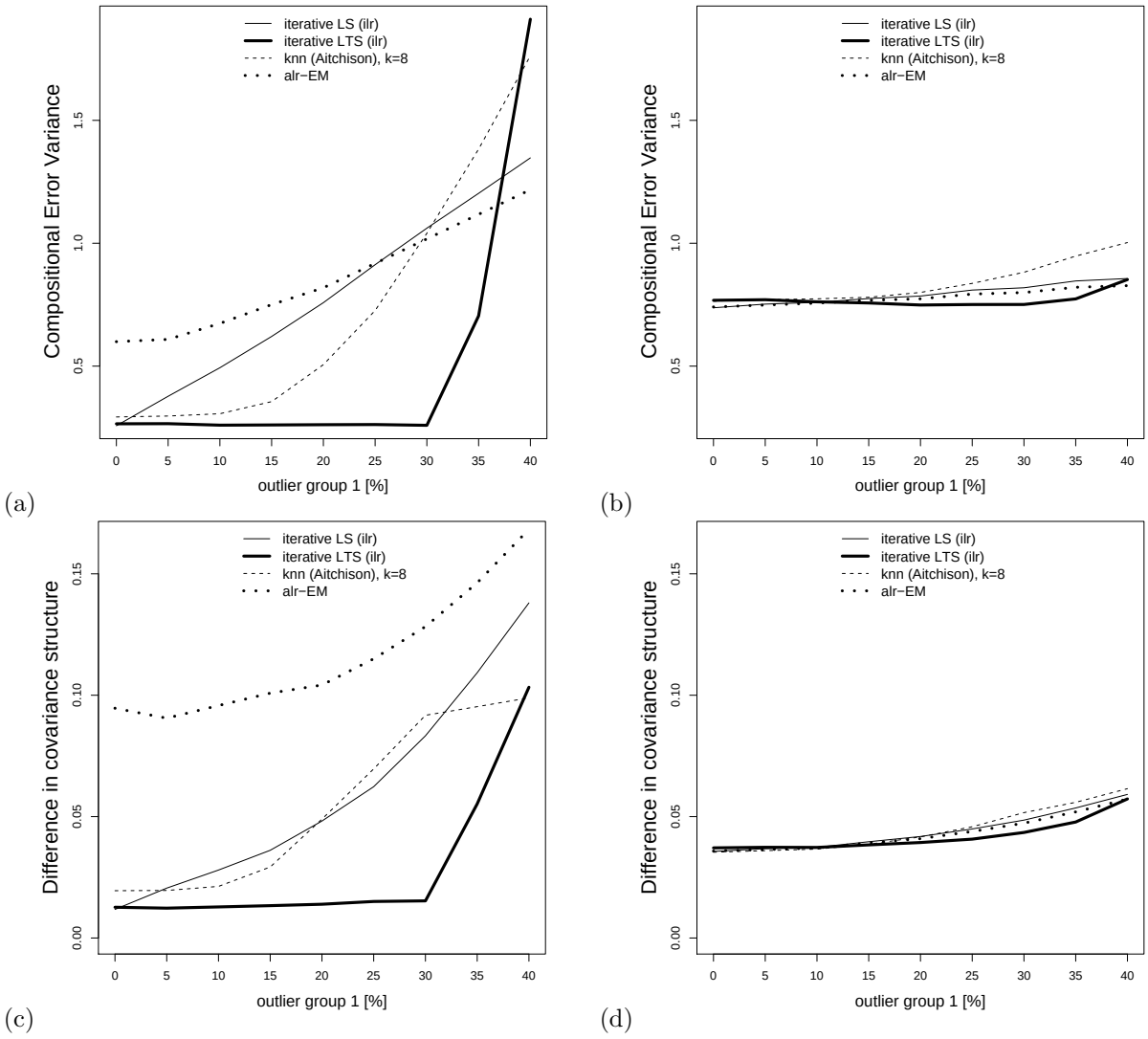

Figure 5: Simulation results (average of the *compositional error variance* and the *difference in variation*) for kNN imputation, model-based imputation using iterative LS and LTS regressions and alr-EM. (a) and (c) shows the results from the second setup, (b) and (d) presents those from the third setup.

either.

We also investigated other imputation methods in the Euclidean and in the Aitchison geometry, such as univariate approaches based on arithmetic and geometric mean, the EM algorithm (Dempster et al., 1977), iterative procedures based on principal component analysis (PCA) (Serneels and Verdonck, 2008) and their robust counterparts (Fritz and Filzmoser, 2008), Bayesian PCA (Oba et al., 2001), and probabilistic PCA (Bishop, 1999). Also various other well-known imputation methods for which an implementation in R (R development core team, 2008) is available were tested (Troyanskaya et al., 2001; Kim et al., 2005; Scholz et al., 2005) (additional methods would be available in R, but sometimes the code is erroneous). All these methods give worse results, and in order to avoid confusion on the plots they are not shown in Figure 4 and 5 . In addition, we tested several variants for kNN imputation (as listed in Section 3.1), in combination with different measures of location (arithmetic mean, median) for the aggregation of the k-nearest neighbors, leading to poorer performance.

6. Conclusions

In general, the estimation of missing values in multivariate data can be done more reliably with multivariate rather than with univariate imputation techniques. However, such procedures are depending on a realistic estimation of the multivariate data structure, because they are either model-based, covariance-based, distance-based, or use a combination of these approaches. The estimation of either the covariance matrix or the distance matrix is sensitive to outliers or data inhomogeneities. Even worse, the underlying geometry of the data plays an essential role for the estimation. In the context of compositional data one has to account for the fact that only relative information in form of ratios between the variables is relevant. In other words, one has to work in the Aitchison geometry rather than in the usual Euclidean geometry (Aitchison, 1986).

We have proposed two imputation methods for estimating missing values in compositional data. The first method uses the information of the k-nearest neighbors, being defined via the Aitchison distance (Aitchison et al., 2000). In principle, this method is not robust against outliers, since an outlying cell in an observation will change the distance. On the other hand, if there are other observations without outlying cells that include valuable information for imputation, the estimation of the missing value might not get much worse. Moreover, one can increase k for kNN imputation in order to collect enough reliable information from the neighbors. This relative robustness of kNN imputation is also visible in the simulation results (see Figure 4 and 5). In presence of contamination the robust adjustment according to Equation (7) led to better results than the adjustment according to Equation (5) in all our tests.

The second proposed method uses kNN imputation to initialize the missing cells. Then regressions are performed in an iterative manner, where in each step one variable serves as the response variable, and the remaining variables as the regressors. This scheme allows to estimate the missing cells in the response with the multivariate information contained in the regressors. However, the regressions have to be done in the appropriate geometry, and therefore we switch to the ilr space (Egozcue et al., 2003) with a special choice of the ilr basis. For homogeneous, outlier-free data one can use least-squares regressions, but in presence of outliers we recommend using robust regressions, like LTS regression (Rousseeuw and Van Driessen, 2006). The numerical study (Section 4) and the simulations (Section 5) have demonstrated that the initial kNN estimation can be essentially improved by this approach. The simulation study has shown that not only the distances between the true and the estimated missings remain very stable, even in presence of outliers, but also the multivariate data structure is well preserved. Our model-based imputation procedure outperformed also other multivariate imputation techniques, even if they are applied in the appropriate geometry.

The simulation was carried out for various different parameter configurations. Three of these simulation setups are presented in this paper whereas one configuration (the third) could be considered as the worst case, especially for our proposed iterative model based imputation technique, because the correlations between the parts are very low, the amount of missing values is equal in each variable, and moderate outliers were generated only in one dimension. Nevertheless, all our results lead to the conclusion that the proposed iterative method has comparable performance to other imputation methods in the worst case, but outperforms existing imputation methods in less extreme situations.

An implementation of our proposed procedures in R (R development core team, 2008) is available in the package *robCompositions* at the Comprehensive R Archive Network, see `http://cran.r-project.org/`.

Acknowledgement

The authors are grateful to the referees for helpful comments and suggestions. This work was supported by the Council of the Czech Government MSM 6198959214.

References

Aitchison, J., 1986. The statistical analysis of compositional data. Chapman & Hall, London. Reprinted in 2003 by Blackburn Press.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Mathematical Geology 32 (3), 271-275.

Beguin, C., Hulliger, B., 2008. The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. Survey Methodology 34 (1), 91-103.

Bishop, C.M., 1999. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B 61, 611–622.

Bren, M., Tolosana-Delgado, R., van den Boogaart, K.G., 2008. News from "compositions", the `R` package. CoDaWork'08, Girona. http://dugi-doc.udg.edu/bitstream/10256/716/1/BREN_cw08_nfc.pdf

Dempster, A.P., Laird, N.M., Rubin D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39, 1-38.

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

Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264, 145-160.

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

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

Filzmoser, P., Hron, K., 2009. Correlation analysis for compositional data. Mathematical Geosciences, in press.

Filzmoser, P., Hron, K., Reimann, C., 2009. PCA for compositional data with outliers. Environmetrics, in press.

Fritz, H., Filzmoser, P., 2008. Plausibility of databases and the relation to imputation methods. VDM Verlag Dr. Müller, Saarbrücken. ISBN: 978-3-8364-5992-1

Kim, H., Golub, G.H., Park, H. 2005. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics. 21 (2), 187-198.

Kovács, L.Ó., Kovács, G.P., Martín-Fernández, J.A., Barceló–Vidal, C., 2006. Major-oxide compositional discrimination in Cenozoic volcanites of Hungary. In Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264, 145-160.

Little, R.J.A., Rubin, D.B., 2002. Statistical analysis with missing data. Wiley, New Jersey, second edition.

Maronna, R., Martin, R.D., Yohai, V.J., 2006. Robust statistics: Theory and methods. John Wiley, New York.

Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology 35 (3), 253-278.

Mateu-Figueras, G., Pawlowsky-Glahn, V., 2008. A critical approach to probability laws in geochemistry. Mathematical Geosciences 40 (5), 489–502.

Oba S., Sato M.A., Takemasa I., Monden M., Matsubara K., Ishii S., 2003. A Bayesian missing value estimation method for gene expression expression profile data. Bioinformatics 19(16), 2088-2096.

Palarea-Albaladejo, J., Martín-Fernández, J. A., 2008. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Computer & Geosciences 34 (8), 902–917.

Pawlowsky-Glahn, V., Egozcue, J.J., 2002. BLU estimators and compositional data. Mathematical Geology 34 (3), 259–274.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, J., 2007. Lecture notes on compositional data analysis. http://hdl.handle.net/10256/297.

Pearson, K., 1897. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London 60, 489–502.

Rousseeuw, P.J., Van Driessen, K., 2006. Computing LTS regression for large data sets. Data Mining Knowl. Disc. 12, 29-45.

R development core team, 2008, R: A language and environment for statistical computing: Vienna, `http://www.r-project.org`.

Schafer, J.L. 1997. Analysis of incomplete multivariate data. Chapman & Hall, London.

Scholz, M., Kaplan, F., Guy, C.L., Kopka, J., Selbig, J., 2005. Non-linear pca: a missing data approach. Bioinformatics 21, 3887-3895.

Serneels, S., Verdonck, T., 2008. Principal component analysis for data containing outliers and missing elements. Computational Statistics & Data Analysis 52 (3), 1712-1727.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520-525.

Van den Boogaart, K.G., Tolosana-Delgado, R., Bren, M., 2006. Concept for handling with zeros and missing values in compositional data. In: Proceedings of IAMG'06 - The XI annual conference of the International Association for Mathematical Geology. University of Liege, Belgium. CD-ROM.

Yucel, R.M., Demirtas, H., 2009. Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. Computational Statistics & Data Analysis. In press. doi:10.1016/j.csda.2009.01.016.