Robust Imputation of Missing Values in Compositional Data Using the R-Package robCompositions

Matthias Templ[*,**], Peter Filzmoser[*], Karel Hron[***]

[*] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria. (templ@tuwien.ac.at)

[**] Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria. (matthias.templ@statistik.gv.at) and

[***] Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Tomkova 40, 779 00 Olomouc, Czech Republic. (hronk@seznan.cz)

Abstract

The aim of this contribution is to show how the R-package `robCompositions` can be applied to estimate missing values in compositional data. Two procedures are summarized, one of them being highly stable also in presence of outlying observations. Measures for information loss are presented, and it is demonstrated how they can be applied. Moreover, we introduce new diagnostic tools that are useful for inspecting the quality of the imputed data.

1 Introduction

1.1 Imputation

Many different methods for imputation have been developed over the last few decades. The techniques for imputation can be subdivided into four categories: univariate methods such as column-wise (conditional) mean or median imputation, distance-based imputation methods such as k-nearest neighbor imputation, covariance-based methods such as the well-known expectation maximization imputation method, and model-based methods such as regression imputation. Most of these methods are able to deal with missing completely at random (MCAR) and missing at random (MAR) missing values mechanism (see, e.g. Little and Rubin [1987]). However, most of the existing methods assume that the data originate from a multivariate normal distribution. This assumption becomes invalid as soon as there are outliers in the data. In that case imputation methods based on robust estimates should be used.

1.2 Compositional Data

Advanced (robust) imputation methods have turned out to work well for data with a direct representation in the Euclidean space. However, this is not the case when dealing with compositional data.

An observation $\mathbf{x} = (x_1, \dots, x_D)$ is called a D-part composition if, and only if, all its components are strictly positive real numbers and all the relevant information is included in the ratios between them [Aitchison, 1986]. One can thus define the *simplex*, which is the sample space of D-part compositions, as

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, \dots, x_D),\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa\} \quad . \tag{1}$$

Note that the constant sum constraint κ implies that D-part compositions are only $D-1$ dimensional, so they are singular by definition. It is, however, possible that the constant κ is different for each observation (for further details, see Hron et˜al. [2008]). In any case, the important property of compositional data is that all information is contained in the ratios of the parts.

The application of standard statistical methods, like correlation analysis or principal component analysis, directly to compositional data can lead to biased and meaningless results [Filzmoser and Hron, 2008a,b]. This is also true for imputation methods [Bren et˜al., 2008, Martín-Fernández et˜al., 2003, Boogaart et˜al., 2006]. A way out is to first transform the data with appropriate transformation methods. Such transformations, preserving the specific geometry of compositional data on the simplex (also called Aitchison geometry), are represented by the family of log-ratio transformations: additive, centered [Aitchison, 1986] and isometric (abbreviated by *ilr*, [Egozcue et˜al., 2003] transformations. Standard statistical methods can then be applied to the transformed data, and the results can be back-transformed.

Compositional data frequently occur in official statistics. Examples are expenditure data, income components in tax data, wage components in the Earnings Structure Survey, components of turnover of enterprises etc., and all data which sum up to a constant or which carry all the information only in the ratios. The problem of missing values in compositional data including outliers is a common problem not only in official statistics, but also in various other fields (see, e.g., [Graf, 2006, Filzmoser and Hron, 2008a]). In the following Section we will briefly review two algorithms for imputation that are described in detail in Hron et˜al. [2008]. Section 3 focuses on the use of the R-package `robCompositions`, and Section 4 introduces some diagnostic tools implemented in this package. The final Section 5 concludes.

2 Proposed Imputation Algorithms

In the following we briefly describe the imputation methods that have been implemented in the R-package `robCompositions`. The detailed description of the algorithms can be found in Hron et~al. [2008].

2.1 k-Nearest Neighbor Imputation

k-nearest neighbor imputation usually uses the Euclidean distance measure. Since compositional data are represented only in the simplex sample space, we have to use a different distance measure, like the Aitchison distance, being defined for two compositions $\mathbf{x} = (x_1, \ldots, x_D)$ and $\mathbf{y} = (y_1, \ldots, y_D)$ as

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

Thus, the Aitchison distance takes care of the property that compositional data include their information only in the ratios between the parts.

Once the k-nearest neighbors to an observation with missing parts have been identified, their information is used to estimate the missings. For reasons of robustness, the estimation is based on using medians rather than means. If the compositional data do not sum up to a constant, it is important to use an adjustment according the sum of all parts prior to imputation. For details, see Hron et~al. [2008].

2.2 Iterative Model-Based Imputation

In the second approach we initialize the missing values with the proposed k-nearest neighbor approach. Then the data are transformed to the $D-1$ dimensional real space using the ilr transformation. Let d_e denote the Euclidean distance. The ilr transformation holds the so-called isometric property,

$$d_a(\mathbf{x}, \mathbf{y}) = d_e(ilr(\mathbf{x}), ilr(\mathbf{y})) \tag{3}$$

[Egozcue and Pawlowsky-Glahn, 2005]. Consequently, one can use standard statistical methods like multiple linear regression, that work correctly in the Euclidean space.

We take a special form of the ilr transformation, namely $ilr(\mathbf{x}) = (z_1, \ldots, z_{D-1})$, with

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}{x_i} \quad \text{for } i = 1, \ldots, D-1 \quad . \tag{4}$$

Here, the compositional part x_1 includes the highest amount of missings, x_2 the next highest, and so on. Thus, when performing a regression of z_1 on

$z_2, \ldots, z_{D-1}$, only z_1 will be influenced by the initialized missings in x_1, but not the remaining ilr variables.

The idea of the procedure is thus to iteratively improve the estimation of the missing values. After the regression of z_1 on $z_2, \ldots, z_{D-1}$, the results are back-transformed to the simplex, and the cells that were originally missing are updated. Next we consider the variable which originally has the second highest amount of missings, and the same regression procedure as before is applied in the ilr space. After each variable containing missings has been proceeded, one can start the whole process again until the estimated missings stabilize. The detailed description of this algorithm can be found in Hron et~al. [2008].

As a regression method we propose to use robust regression, like LTS regression (see Maronna et~al. [2006]), especially if outliers might be present in the data.

3 Using the R-package `robCompositions` for Imputing Missing Values

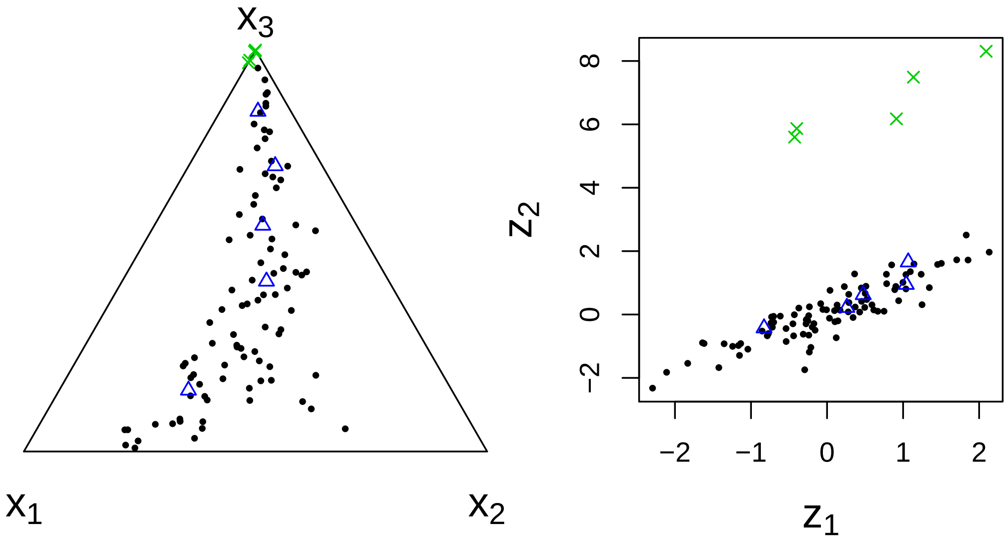
3.1 Data

The package includes the three compositional data sets *aitchison359*, *aitchison360*, and *aitchison395*, that have been published in Aitchison [1986]. In the following, however, we will use simulated data, where the data structure and outliers are exactly known. The data generation is the same as described in Hron et~al. [2008], and a plot of the data set in shown in Figure 1 for the original data (left) and for the ilr transformed data (right): We took 90 observations with 3 parts that are normally distributed on the simplex (i.e. they are multivariate normally distributed in the 2-dimensional ilr space). A group of 5 outliers (*group 1*) is added (green crosses in Figure 1) that are potential outliers in the Aitchison and in the Euclidean space. Another group (*group 2*) of 5 outliers (blue triangles in Figure 1) is only affecting the Euclidean space. Note that both types of outliers are simulated to have a considerably higher sum of their parts, which is not visible in the ternary diagram [Aitchison, 1986] in Figure 1 (left) where the parts are re-scaled to have sum 1.

The generated (complete) data are stored in the list element `z2` of object x. Among the non-outliers we set 20% of the values in the first part, 10% in the second, and 5% in the third part to missing, using an MCAR mechanism. The new data set is stored in the list element `zmiss` of object x.

3.2 Usage of the Imputation Methods Within the Package

We apply k-nearest neighbor imputation for the generated compositional data set, and use the parameter $k = 6$:



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

```
> library(robCompositions)
> xImp <- impKNNa(x$zmiss, k=6)
```

As a default, Aitchison distances are used for identifying the k-nearest neighbors (further options are provided, see help file). By default, the median is taken for re-scaling the k-nearest neighbors for imputation, but also other choices are possible.

The resulting object `xImp` is of class

```
> class(xImp)
```

```
[1] "imp"
```

A print, a summary, and a plot method are provided for objects of this class:

```
> methods(class = "imp")
```

```
[1] plot.imp*    print.imp*   summary.imp*

   Non-visible functions are asterisked
```

```
> xImp
```

```
 ----------------------------------------
[1] "31 missing vales were imputed"
 ----------------------------------------
```

Various informations are included in the object `xImp`, which can be accessed easily:

```
names(xImp)

[1] "xOrig"    "xImp"     "criteria" "iter"     "w"
"wind" "metric"
```

The list element `xOrig` contains the original data, `xImp` is the imputed data set, `w` contains the number of missing values, and `wind` includes the indices of the missing values (imputed values). All this information is needed in order to provide suitable summaries and diagnostic plots.

The iterative model-based imputation method is applied with:

```
> xImp1 <- impCoda(x$zmiss, method='lm')
> xImp2 <- impCoda(x$zmiss, method='ltsReg')
```

The first command uses classical least-squares regression within the algorithm, the second command takes robust LTS regression.

4 Information Loss, Uncertainty, and Diagnostics

The quality of the imputed values can be judged by different criteria. We can use information loss criteria and compute the differences of the imputed to the observed data. If the observed data are known, we can use the bootstrap technique for measuring the uncertainty of the imputation. If the observed data are not known, diagnostic plots can be used for visualizing the imputed values.

4.1 Information Loss Measures

We compare the imputed and the original data values by two different criteria:

Relative Aitchison distance: **(RDA)** Let $M \subset \{1, \ldots, n\}$ denote the index set referring to observations that include at least one missing cell, and $n_M = |M|$ be the number of such observations. We define the *relative Aitchison distance* as

$$\frac{1}{n_M} \sum_{i \in M} d_A(\mathbf{x}_i, \hat{\mathbf{x}}_i) \tag{5}$$

where $\mathbf{x}_i$ denotes the original composition (before setting cells to missing), and $\hat{\mathbf{x}}_i$ denotes the composition where only the missing cells are imputed.

Difference in variations: **(DV)** We use the variation matrix $\mathbf{T} = [t_{ij}]$, with

$$t_{ij} = \mathrm{var}\left(\ln \frac{x_i}{x_j}\right),\ i, j = 1, \ldots, D\,,$$

and the empirical variance for var. Thus, t_{ij} represents the variance of the log-ratio of the parts i and j. Here, only the non-outlying original observations are considered for computing $\mathbf{T}$. On the other hand, $\tilde{\mathbf{T}} = [\tilde{t}_{ij}]$ denotes the variation matrix computed for the same observations, where all missing cells have been imputed. Then we define the *difference in variations* as

$$\frac{2}{D(D-1)} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} |t_{ij} - \tilde{t}_{ij}| \qquad (6)$$

Thus, RDA measures closeness of the imputed values in the Aitchison geometry, whereas the influence of the imputation to the multivariate data structure is expressed by DV.

Using the iterative model-based algorithm for our test data set, we can show that the robust procedure based on LTS regression gives more reasonable results than its classical counterpart (the code for computing the measures is snipped):

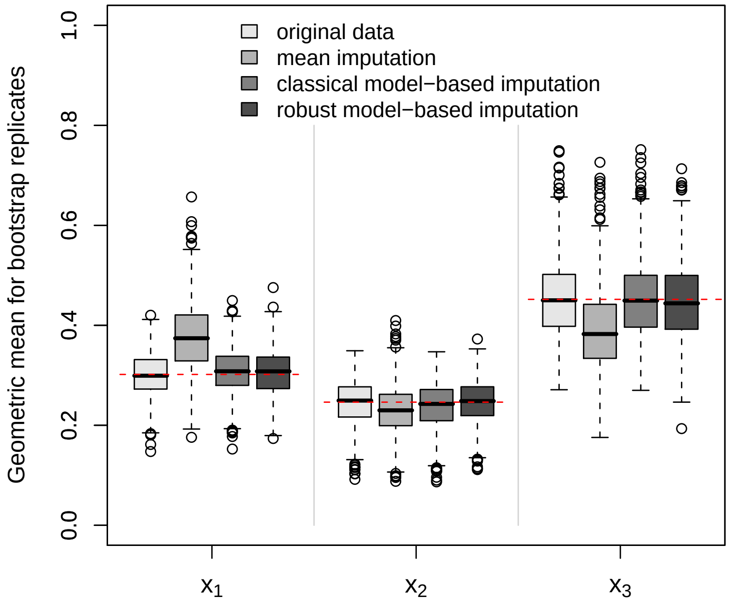
```
[1] "RDA: iterative lm approach: 0.533"

[1] "RDA: iterative ltsReg approach: 0.317"

[1] "DV: iterative lm approach: 0.052"

[1] "DV: iterative ltsReg approach: 0.006"
```

4.2 Measuring the Uncertainty of the Imputations

Little and Rubin [1987] suggests to estimate standard errors for estimators via bootstrapping, and he outlines two approaches - a modified bootstrap approach and a modified jackknife procedure - to measure consistent standard errors when data will be imputed.

We draw bootstrap samples from both, the original data without missings, and the data where some values were set to missing, hereby using the same random seeds. For the latter bootstrap samples we impute the missing values with mean imputation (column-wise arithmetic mean), and classical and robust iterative model-based imputation. We are interested in the geometric mean of each variable. Figure 2 shows boxplots of the resulting geometric means (computed only for the non-outlying observations) for $r = 1000$ bootstrap replicates. The red horizontal lines indicate the geometric means for the original data without outliers.

It is clearly visible that mean imputation - a simple method still frequently applied - can lead to higher uncertainty, and that the results are biased. The model-based procedures have a very similar behavior as the original data.



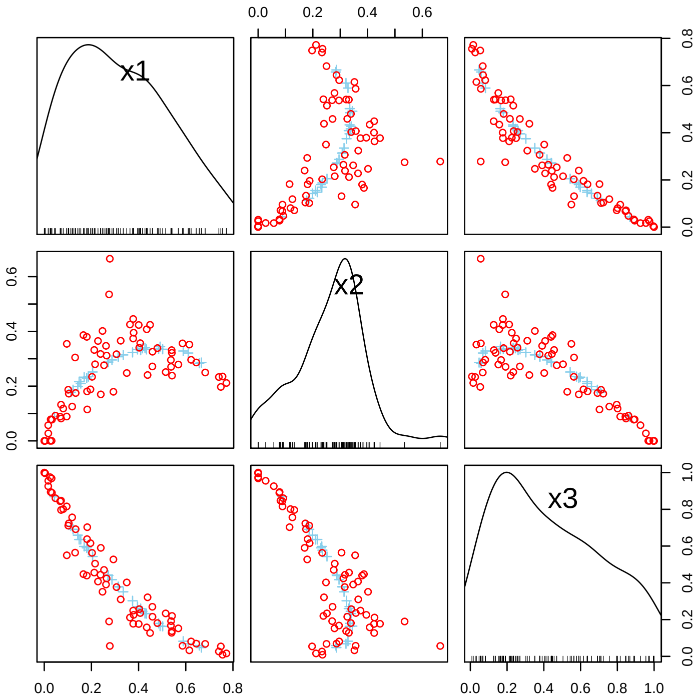
Figure˜2: Boxplot comparison of the estimated column-wise geometric means of 1000 bootstrap replicates of the original data set and the data sets where missing values were imputated with different methods. The red line is the column-wise geometric mean of the original data. Outliers are excluded in the computation of the geometric means.

4.3 Diagnostic Plots

Here we do not assume knowledge about the observed values. The goal is to visualize the imputed values in an appropriate way. Because of space limitations we only show results for the robust model-based procedure.

The first diagnostic plot is a multiple scatterplot where the imputed values are highlighted. The plot can be generated with `plot(xImp2, which=1)`. Figure 3 shows the result, and we can see that the imputed values are placed on the regression hyperplane(s). If this should be avoided, one can add random noise to the imputed values. This can be done with the function `impCoda()` using the parameter `method` = `ltsReg2` which considers the standard deviation of the residuals for generating the random noise. This plot can also be generated for instance for log-ratio transformed data [Aitchison, 1986], taking care of the compositional nature of the data.

A parallel coordinate plot [Wegman, 1990] can be generated by `plot(xImp2, which=2)` (plot not shown here). The imputed values in certain variables are



Figure~3: Multiple Scatterplot to highlight imputed observations.

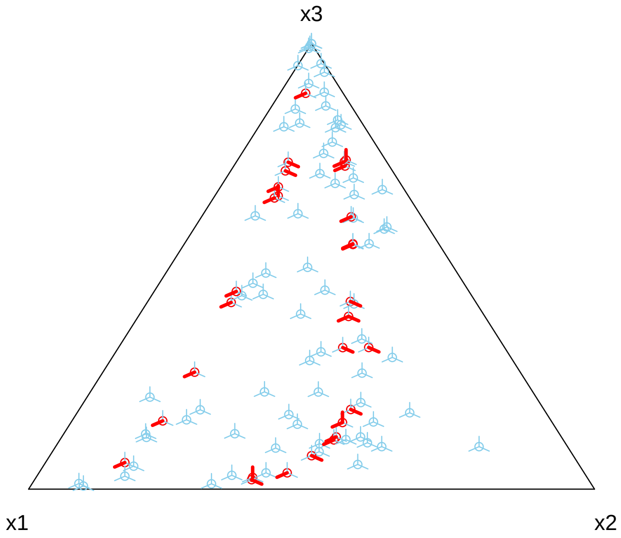
highlighted. One can select variables interactively, and imputed values in any of the selected variables will be highlighted.

The third diagnostic plot (see Figure 4), a ternary diagram [Aitchison, 1986], can be generated by `plot(xImp2, which=3, seg1=FALSE)`. The 3-part compositions are presented by three spikes, pointing in the directions of the corresponding three variables. The spikes of the imputed values are highlighted. This presentation allows gaining a multivariate view of the data, being helpful for interpreting possible irregularities of imputed values.

5 Conclusions

We provide the R-package `robCompositions` which includes advanced methods for imputation for compositional data. We have shown how the imputation methods described in Hron et~al. [2008] can be applied with the package. The methods are especially designed for data including outliers. The performance of the methods is outlined in the original paper.

The package includes possibilities for evaluating the quality of the imputed values: One can compute measures for information loss, use bootstrapping



Figure˜4: Ternary diagram with special plotting symbols for highlighting imputed parts of the compositional data.

for estimating bias and uncertainty of parameters, and visualize the imputed values with diagnostic tools. For 3-dimensional compositions the proposed ternary plot is designed to highlight how well imputations are made and which compositions were imputed. Note that many additional options can be used within these plots given by the arguments of the plotting function.

References

J.˜Aitchison. *The Statistical Analysis of Compositional Data.* Wiley, New York, 1986.

K.G. Boogaart, R.˜Tolosana-Delgado, and M.˜Bren. Concept for handling with zeros and missing values in compositional data. In E.˜Pirard, editor, *CProceedings of IAMG'06 - The XI annual conference of the International Association for Mathematical Geology.* University of Liege, Belgium. CD-ROM., 2006. 4 pages.

M.˜Bren, R.˜Tolosana-Delgado, and K.G. van˜den Boogaart. News from compositions, the R package. In D.˜Estadella, J.˜Martín-Fernández, and J.˜Antoni, editors, *CoDaWork'08.* Universitat de Girona. Departament d'Informática i Matemática Aplicada, 2008.

J.J. Egozcue and Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828, 2005. doi: 10.1007/s11004-005-7381-9. URL http://www.springerlink.com/content/fx037244708561v5/.

J.J. Egozcue, V.~Pawlowsky-Glahn, G.~Mateu-Figueras, and C.~Barceló-Vidal. Isometric log-ratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003. doi: 10.1023/A:1023818214614. URL http://www.springerlink.com/content/wx1166n56n685v82/.

P.~Filzmoser and K.~Hron. Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40(3):233–248, 2008a. doi: http://dx.doi.org/10.1007/s11004-007-9141-5. URL http://www.springerlink.com/content/d662421553216861.

P.~Filzmoser and K.~Hron. Correlation analysis for compositional data. Research report sm-2008-2, Department of Statistics and Probability Theory, Vienna University of Technology, 2008b. URL http://www.statistik.tuwien.ac.at/forschung/SM/SM-2008-2complete.pdf.

M.~Graf. Swiss earnings structure survey. compositional data in a stratified two-stage sample. Metholology report, isbn: 3-303-00338-6, Swiss Federal Statistical Office, 2006. URL http://www.bfs.admin.ch/bfs/portal/de/index/themen/00/07/blank/02.Document.77975.pdf.

K.~Hron, M.~Templ, and P.~Filzmoser. Imputation of compositional data using robust methods. Research report sm-2008-4, Department of Statistics and Probability Theory, Vienna University of Technology, 2008. URL http://www.statistik.tuwien.ac.at/forschung/SM/SM-2008-4complete.pdf.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, New York, 2006.

J.~Martín-Fernández, C.~Barceló-Vidal, and V.~Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003. doi: 10.1023/A:1023866030544. URL http://www.springerlink.com/content/ku816485q4264772/.

E.J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85:664–675, 1990.