Classical and robust imputation of missing values for compositional data using balances

K. HRON[1], M. TEMPL[23], and P. FILZMOSER[2]

[1]Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
hronk@seznam.cz
[2]Department of Statistics and Probability Theory - Vienna University of Technology, Austria
[3]Department of Methodology - Statistics Austria, Austria

Most statistical methods cannot be directly applied to data sets including missing observations. For this reason, many different methods for imputation of multivariate data have been developed over the last few decades. As an example, 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. However, one usually assumes that the data originate from a multivariate normal distribution, which is no longer valid in the presence of outliers in the data.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. 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. This, however, is not the case for compositional data with the Aitchison geometry, and thus a different approach for imputation has to be used.

In the contribution, two methods for imputation of compositional data will be presented (Hron et al., 2010). In the first proposed method, based on $k$nn imputation, the Aitchison distance is used to search for 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. 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. Note that this was not necessary for finding the k-nearest neighbors, because the Aitchison distance is the same for any D-part compositions $\mathbf{x}$ and $\mathbf{y}$ belonging to equivalence classes $\underline{\mathbf{x}} = \{c\mathbf{x},\, c \in \mathbf{R}^+\}$ and $\underline{\mathbf{y}} = \{c\mathbf{y},\, c \in \mathbf{R}^+\}$. 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. 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 (preferably robust) regression, but we have to work in coordinates. For this purpose we choose the isometric logratio transformation, and concretely balances that make a meaningful interpretation possible. However, already for constructing the balances a data matrix with complete information is needed. 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. Thus, the choice of the balances 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 $D-2$ balances. Thus, using such a sequential binary partition will cause that as few as possible balances are affected by the missing values.

The advantages of both methods will be illustrated on a simulation study and a real data example.

References

Hron, K., M. Templ, and P. Filzmoser (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics and Data Analysis 54*(12), 3095–3107.