Chemometrics and Intelligent Laboratory Systems 155 (2016) 183–190

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Imputation of rounded zeros for high-dimensional compositional data

Matthias Templ [a], Karel Hron [b], Peter Filzmoser [a], Alžběta Gardlo [b,c]

[a] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstrasse 8-10, Vienna A-1040, Austria
[b] Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, Olomouc, CZ 77146, Czech Republic
[c] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University, Hněvotínská 5, Olomouc CZ77900, Czech Republic

ARTICLE INFO

Article history:
Received 12 February 2016
Received in revised form 14 April 2016
Accepted 18 April 2016
Available online 27 April 2016

Keywords:
High-dimensional compositional data
Rounded zeros
Imputation

ABSTRACT

High-dimensional compositional data, multivariate observations carrying relative information, frequently contain values below a detection limit (rounded zeros). We introduce new model-based procedures for replacing these values with reasonable numbers, so that the completed data set is ready for use with statistical analysis methods that rely on complete data, such as regression or classification with high-dimensional explanatory variables. The procedures respect the geometry of compositional data and can be considered as alternatives to existing methods. Simulations show that especially in high-dimensions, the proposed methods outperform existing methods. Moreover, even for a large number of rounded zeros, the new methods lead to an improved quality of the data, which is important for further analyses. The usefulness of the procedure is demonstrated using a data example from metabolomics.

1. Introduction

High-dimensional data refer to a situation with many variables, often many more variables than observations. Such data are often connected to genetics (as microarray data), but they also arise in other fields that integrate chemometrical and biological processes, such as in proteomics or metabolomics, as well as in other natural sciences. In biological applications, due to high-tech measurement devices, it is possible for each observation to produce hundreds or even thousands of variables (potential biomarkers) that should be further analyzed statistically. However, frequently there are not enough samples at hand, such as when a rare metabolomic disease is of primary interest. Consequently, analysis of this kind of data requires specific statistical methods that can cope with the situation of having more variables than observations. Examples are procedures with various inference goals (pattern recognition, calibration, clustering). They include methods based on singular value decomposition or distance-based methods, like partial least squares (PLS) regression, Lasso regression, or hierarchical clustering [1].

An additional problem arises when the absolute values of the variables (parts) are not the primary interest but rather their relative values on a whole (which we refer to as compositional data, or compositions for short) [2,3]. This means that any possible rescaling of the data does not alter the relevant information contained in ratios between the parts. As a particular case, a representation of the data in proportions should not alter the results of a meaningful statistical analysis. These properties are fulfilled by expressing the compositions in orthonormal coordinates with respect to their specific geometry, the Aitchison geometry on the simplex [4]. The above two features, i.e. high dimensionality and compositional (or mixture) nature of data, are also commonly shared by different types of high-dimensional data in chemometrics; the corresponding methods for compositional data were recently successfully applied in metabolomics [5,6].

Unfortunately, none of the above-mentioned statistical methods are able to process data that contain measurement artefacts such as missing values (pure absence of the measurement in some entries) or values below a detection limit (resulting as effect of rounding errors; we also refer to rounded zeros). Especially values below a detection limit frequently occur in natural sciences related to data from chemometrics or from geochemistry. In metabolomics, for example, rounded zeros usually arise from the preprocessing step, when values below a certain threshold are set to zero in order to suppress possible effects of inaccuracy of the measurement device. For this reason, we consider the zeros to be the result of a rounding error rather than of a pure absence of the molecule in the concrete variable. Consequently, a proper imputation of rounded zeros must precede any further statistical analysis. Although in case of standard multivariate data a comprehensive methodology exists [7], even applicable to the imputation of rounded zeros in high-dimensional data [8,9], it fails in case of compositional data. Due to their specific nature, each value to be imputed needs to be considered in a relative sense, as ratios with the other parts in a composition. Methods for the imputation of values below a detection limit are already

E-mail addresses: templ@statistik.tuwien.ac.at (M. Templ), hronk@seznam.cz (K. Hron), p.filzmoser@tuwien.ac.at (P. Filzmoser), alzbetagardlo@gmail.com (A.ě Gardlo).

http://dx.doi.org/10.1016/j.chemolab.2016.04.011
0169-7439/

developed and extensively used in practice [10–14]. However, most of these methods fail in case of high-dimensional compositional data sets. Finally, it is worth noting that in many cases, missing values are mixed up with rounded zeros in chemometrics. Rounded zeros are non-detects that result either from rounding effects (rounding to zero) or as values below a certain detection limit (censored values), but it is rather common in statistics to denote by missing values unobserved uncensored entries.

The paper is structured as follows: Section 2 reviews the construction of isometric log-ratio coordinates, as well as available methods for the imputation of rounded zeros in compositional data. A new method using a model-based procedure based on partial least squares regression is proposed in Section 3, and Section 4 provides a modification of this method using variable selection. Advantages and shortcomings of the approaches are analyzed using simulated data and a data set from metabolomics (Section 5). Finally, the main features of the presented methods and the comparison to existing methodology are summarized in Section 6.

2. Available methods for rounded zeros imputation

The replacement of rounded zeros represents a constrained version of missing values imputation. Namely, when x_{ij} represents a rounded zero for a particular observation i and a variable j, it holds that $x_{ij}<t_{ij}$, where t_{ij} is a threshold, i.e., typically the detection limit. For the purpose of imputation, a regression-based algorithm was proposed in [14] which, however, does not work in high-dimensional situations. The initialization of the iterative procedure is done by assigning 2/3 of the detection limit to each of the affected data entries [10]. Note that for more than approximately 10% rounded zeros, this might result in a serious distortion of the multivariate data structure, even new outlying observations might arise. Thus, a substantial improvement of the initial imputation is necessary. The crucial point is to express the threshold values in log-ratio coordinates [12,14]. This guarantees that the estimated values are placed below the detection limit throughout the estimation process.

2.1. Isometric log-ratio (ilr) coordinates

Because the new methods for rounded zero imputation proposed later on in this paper are built on isometric log-ratio coordinates [15], we provide more background on compositional data analysis in the following.

The relative scale of compositional data as well as their inherent principles like scale invariance are reflected by the Aitchison geometry on the simplex, the sample space of representations of compositional data [3]. The Aitchison geometry has properties of any Euclidean vector space, thus it seems to be intuitive to find a proper orthonormal basis with respect to this geometry and to express the compositions in the corresponding coordinates, we refer to the isometric log-ratio (ilr) coordinates (see [15]). Indeed, such a representation enables us to express compositional data in the usual Euclidean real space, for which most standard statistical methods are designed [16]. Let $\mathbf{X}=(\mathbf{x}_1,\dots,\mathbf{x}_D)$ denote a compositional data matrix with n observations as rows and D compositional parts as columns. A re-ordered composition with the l-th part, $l=1,\dots,D$, moved to the first position is denoted by $\mathbf{X}^{(l)}=(\mathbf{x}_l,\mathbf{x}_1,\dots,\mathbf{x}_{l-1},\mathbf{x}_{l+1},\dots,\mathbf{x}_D)=(\mathbf{x}_1^{(l)},\mathbf{x}_2^{(l)},\dots,\mathbf{x}_l^{(l)},\mathbf{x}_{l+1}^{(l)},\dots,\mathbf{x}_D^{(l)})$.

A particular choice of orthonormal coordinates leads to $D-1$ new coordinates $\mathbf{Z}^{(l)}=(\mathbf{z}_1^{(l)},\dots,\mathbf{z}_{D-1}^{(l)})$,

$$\mathbf{z}_j^{(l)}=\sqrt{\frac{D-j}{D-j+1}}\ \ln\frac{\mathbf{x}_j^{(l)}}{\sqrt[D-j]{\prod_{k=j+1}^{D}\mathbf{x}_k^{(l)}}},\ j=1,\dots,D-1; \tag{1}$$

division of vectors, columns of the compositional data matrix, is performed element-wise. The inverse mapping of $\mathbf{Z}^{(l)}$ to the original (permuted) composition $\mathbf{X}^{(l)}$ is then given by

$$\begin{aligned}
\mathbf{x}_1^{(l)} &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}}\mathbf{z}_1^{(l)}\right),\\
\mathbf{x}_j^{(l)} &= \exp\left(-\sum_{k=1}^{j-1}\frac{1}{\sqrt{(D-k+1)(D-k)}}\mathbf{z}_k^{(l)}+\frac{\sqrt{D-j}}{\sqrt{D-j+1}}\mathbf{z}_j^{(l)}\right),\\
\mathbf{x}_D^{(l)} &= \exp\left(-\sum_{j=1}^{D-1}\frac{1}{\sqrt{(D-j+1)(D-j)}}\mathbf{z}_j^{(l)}\right),
\end{aligned} \tag{2}$$

$j=2,\dots,D-1$. Consequently, the obtained compositions can be represented as vectors with a constant sum constraint, such as proportions or percentages. Due to the dimension of the Aitchison geometry (equal to $D-1$) that reflects the nature of compositional data, it is not possible to assign a coordinate to each of the original compositional parts simultaneously. However, from its construction, the first coordinate $\mathbf{z}_1^{(l)}$ includes all relative information of the part $\mathbf{x}_l$ to the remaining parts, and can thus be interpreted in terms of dominance of part $\mathbf{x}_l$ with respect to the other parts in the composition. Since this part $\mathbf{x}_l$ does not occur in the other coordinates $\mathbf{z}_2^{(l)},\dots,\mathbf{z}_{D-1}^{(l)}$, these coordinates explain the remaining log-ratios in the composition [17]. This choice of coordinates is of particular importance in the imputation context [18,14] as well as in other applications (see, e.g., [19,6]).

2.2. Existing methodology

The available methods for rounded zeros imputation are collected in the R-package *zCompositions* [20]. We will briefly review these methods, and employ them in the experimental part of the paper.

2.2.1. Multiplicative replacement (mult repl)

This method imputes left-censored compositional values by a given fraction of the corresponding detection limit. The default fraction is 2/3 times the detection limit of a variable. Multiplicative adjustment is applied in such a manner that the row-wise sums are made equal to the original values including rounded zeros whenever the data are in closed form, i.e. if they have to sum up to a constant. In this case, the absolute values are not preserved. Multiplicative replacement does not modify the original values above the detection limit if the data are not presented in a closed form.

2.2.2. Multiplicative log-normal replacement (mult lognorm)

[21] consider the univariate log-odds for the i-th variable (for values above detection limit). They model the compositions using a multiplicative logistic normal mixture for this purpose.

2.2.3. Multiplicative Kaplan–Meier smoothing spline replacement (mult KMSS)

This method replaces left-censored rounded zeros by averaging (geometric mean) random draws from a cubic smoothing spline fit. This spline is fit to the inverse Kaplan–Meier empirical cumulative distribution function to values below the corresponding limit of detection or censoring threshold, and the values below detection are replaced by the fitted values. Note that this method works in a univariate manner, applied independently to each compositional part containing values below detection. However, afterwards multiplicative adjustment is applied to preserve the multivariate compositional properties of the