Zero Replacement in Compositional Data Sets

J. A. Martín-Fernández[1], C. Barceló-Vidal[1], and V. Pawlowsky-Glahn[2]

[1] Dept. Informàtica i Matemàtica Aplicada, Universitat de Girona, Campus Montilivi – Edifici PI, E-17071 Girona, Spain. (e-mail: jamf@ima.udg.es)
[2] Dept. Matemàtica Aplicada III, ETSECCPB, Universitat Politècnica de Catalunya, E-08034 Barcelona, Spain. (e-mail: vera.pawlowsky@upc.es)

Abstract. The sample space of compositional data is the open simplex. Therefore, zeros in a compositional data set are identified either with below detection limit values, or lead to a division of the data set into different subpopulations with the corresponding lower dimensional sample space. Most multivariate data analysis techniques require complete data matrices, thus calling for a strategy of imputation of zeros in the first case. Existing replacement methods of rounded zeros are reviewed, and a new method is proposed, who's properties are analyzed and illustrated. The method is applied in a hierarchical cluster analysis of compositional data.

1 Introduction

Compositional data are by definition proportions of some whole. Thus, their natural sample space is the open simplex and interest lies in the relative behavior of the components. The open simplex is defined as (Aitchison, 1986)

$$\mathcal{S}^D = \{(x_1, x_2, \ldots, x_D)' : x_j > 0; j = 1, 2, \ldots, D; x_1 + x_2 + \cdots + x_D = 1\}.$$

Any vector of positive components, $\mathbf{y} \in \Re^D_+$, can be projected into the simplex by the *closure operation* $\mathcal{C}(\mathbf{y}) = (y_1/\sum y_j, y_2/\sum y_j, \ldots, y_D/\sum y_j)'$. The only operations known to induce a vector space structure on the simplex are the *perturbation operation*, $\mathbf{p} \circ \mathbf{x} = \mathcal{C}(p_1 x_1, p_2 x_2, \ldots, p_D x_D)'$, defined on $\mathcal{S}^D \times \mathcal{S}^D$, and the *power transformation*, $\alpha \diamond \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)'$, defined on $\Re \times \mathcal{S}^D$. Perturbation can be proven to be equivalent to translation in $\Re^D$ and power transformation to the scalar product using the centered-logratio transformation (clr). This transformation has been defined by Aitchison (1986) as $\mathrm{clr}(\mathbf{x}) = (\ln(x_1/g), \ln(x_2/g), \ldots, \ln(x_D/g))'$, with $g = (\prod_{i=1}^D x_i)^{1/D}$. Another operation on the simplex, analogous to projection onto a smaller dimensional space in $\Re^D$, is obtained through the concept of *subcomposition.* It is defined as $\mathbf{x}_S = \mathcal{C}(\mathbf{S}\mathbf{x})$, $\mathbf{x} \in \mathcal{S}^D$, where $\mathbf{S}$ is a $(S \times D)$ selecting matrix with all elements zero except one in each row and at the most one in each column. The subcomposition $\mathbf{x}_S$ belongs to the simplex $\mathcal{S}^S$. A distance, compatible with the previously defined operations, is Aitchison's distance $d_a(\mathbf{x}, \mathbf{x}^*) = d_e(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{x}^*))$, where $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$ and d_e is the Euclidean distance. Its properties have been discussed in Martín-Fernández et al. (1998) and in Aitchison et al. (1999).

The “Achilles heel” of $d_a(\mathbf{x}, \mathbf{x}^*)$ is the presence of zero values in the data, as it is not possible to take the logarithm of zero. Zero values are present in many applications as, for example, in a household budget spending nothing on the commodity group “tobacco and alcohol”, or in a rock specimen containing “no trace” of a particular mineral. In compositional data we distinguish two kinds of zeros: *essential zeros* and *rounded zeros.* The zero in the household spending pattern is essential. The zero in a particular mineral is usually a rounded zero, *i. e.* it indicates that no quantifiable proportion of the mineral can be recorded according to the accuracy of the measurement process. In hierarchical cluster analysis the presence of an essential zero in a component is an indication that the observation belongs to a different group and straightforward division of the sample is advisable. The principal problem refers to rounded zeros.

The purpose of this paper is to revise, from a theoretical point of view, the additive method of replacement suggested by Aitchison (1986), whose drawbacks have been described from an empirical point of view in Tauber (1999), and to propose a new method of replacement of zeros in compositional data. First, we present the solution proposed by Aitchison (1986). Next, we summarize the most usual non-parametric approaches for missing values with non-compositional data and then we propose a new method of zero replacement. Finally, we present an example where the proposed method is applied to a compositional data set.

2 Additive replacement strategy

Aitchison (1986) suggests that an observation $\mathbf{x} \in \mathcal{S}^D$ containing C rounded zeros can be replaced by a new observation $\mathbf{r} \in \mathcal{S}^D$ without zeros according to the following replacement rule:

$$r_j = \begin{cases} \frac{\delta(C+1)(D-C)}{D^2}, & \text{if } x_j = 0 \ , \\ \\ x_j - \frac{\delta(C+1)C}{D^2}, & \text{if } x_j > 0 \ , \end{cases} \tag{1}$$

where δ is smaller than a given threshold derived from the measurement process.

Note that the constant-sum-constraint of compositional data forces to modify both the zero and the non-zero values. Moreover, the imputed value r_j depends not only on δ but also on the dimension D and the number C of zeros. Note also that a different δ_j could be considered for every component x_j leading to a slightly more complicated expression.

Due to the fact that the transformation (1) of non-zero values is additive, it holds that $r_k/r_l \neq x_k/x_l$, for x_k, x_l non-zero values, and the value of the new ratios r_k/r_l depends on δ. Therefore, Aitchison’s distance between two replaced observations is extremely sensitive to changes in δ as illustrated empirically by Tauber (1999).

3 Replacement strategies for non-compositional data

Let $\mathbf{Y}$ be a data set with missing values in real space $\Re^D$. If the goal is to perform a cluster analysis based on a hierarchical clustering method using the Euclidean distance, it is necessary to complete first the matrix of distances between observations. Several strategies have been suggested in the literature for that purpose. The one by Krzanowski (1988) can be synthesized as follows: (i) omit any variable that has a missing value when computing the distance between two observations and work only with those variables that have all values present for both the observations concerned; (ii) if the previous step means working with S variables instead of D, inflate the resulting distance by a factor D/S. To see that this strategy is not suitable for compositional data, consider the following example: given three compositional observations $\mathbf{x} = (0, 0.8, 0.2)$, $\mathbf{x}^* = (0.95, 0.04, 0.01)$, and $\mathbf{x}' = (0.06, 0.76, 0.18)$, the strategy of Krzanowski implies to consider the subcompositions formed by the second and third variables: $\mathbf{x_S} = (0.8, 0.2)$, $\mathbf{x_S}^* = (0.8, 0.2)$, and $\mathbf{x_S}' = (0.81, 0.19)$. Assuming that the zero in sample $\mathbf{x}$ is actually a very small positive value, we expect $\mathbf{x}$ and $\mathbf{x}'$ to be more similar than $\mathbf{x}$ and $\mathbf{x}^*$. Nevertheless, we obtain that $d_a(\mathbf{x_S}, \mathbf{x_S}^*) = 0$ and $d_a(\mathbf{x_S}, \mathbf{x_S}') = 0.07$.

The most common strategy to complete the matrix of distances is to employ "imputation", *i. e.* the insertion of an estimate for each missing value, thereby completing the data set, and then calculate the matrix of distances. When the missing values are actually censored data, that is, when the values for some variables are reported as "less than" a given threshold value, a simple imputation can be considered. For a "small" proportion of "less than" values (not more than 10%) a simple-substitution method using 0.55 of the threshold value is suggested in Sandford et al. (1993). More general imputation methods are exposed in Little and Rubin (1987).

All these imputation methods have one thing in common: the canonical projection $\Pi(\mathbf{y})$ on the non-missing variables of observation $\mathbf{y}$ is identical to the same projection $\Pi(\mathbf{z})$ of the replaced observation $\mathbf{z}$. Also, if $\mathbf{y}$ and $\mathbf{y}^*$ have "common" missing values, *i. e.* missing values on the same variables, it holds that $y_j - y_j^* = z_j - z_j^*$ for y_j, y_j^* non-missing values and z_j, z_j^* the corresponding replacement. Furthermore, if the imputation method assigns the same replacement value to every missing component y_j of the two observations, then $d_e(\mathbf{z}, \mathbf{z}^*)$ does not depend on the imputated values and it is identical to the Euclidean distance between the projections $d_e(\Pi(\mathbf{y}), \Pi(\mathbf{y}^*))$. With the above features in mind, let us proceed to define a suitable replacement method for zeros in compositional data.

4 Multiplicative replacement strategy

Let be $\mathbf{x} \in \mathcal{S}^D$ and assume it has C zeros. We propose to replace $\mathbf{x}$ with an observation $\mathbf{r} \in \mathcal{S}^D$ without zeros using the expression

$$r_j = \begin{cases} \delta_j, & \text{if } x_j = 0, \\ x_j(1 - \sum_{k|x_k=0} \delta_k), & \text{if } x_j > 0, \end{cases} \tag{2}$$

where δ_j is the imputated value on the component x_j. Following Sandford et al. (1993), whenever δ_j is equal to 0.55 of the threshold determined from the measurement process corresponding to component x_j, a simple-substitution in the simplex is obtained.

The multiplicative modification of non-zero values in (2) has the following desirable properties not satisfied by (1):

1. It is "natural" in the sense that, if the imputated values δ_j in an observation $\mathbf{x}$ are equal to the "true" censored values, then $\mathbf{r}$ recovers the "true" observation.
2. It is coherent with the basic operations in the simplex, *i. e.* if a selecting matrix $\mathbf{S}$ of non-zero components of observation $\mathbf{x}$ is considered, and $\mathbf{x}_S = \mathcal{C}(\mathbf{Sx})$ is the subcomposition obtained, denoting by $\mathbf{r}_S = \mathcal{C}(\mathbf{Sr})$ the subcomposition derived from the replacement vector, the following properties hold:
   (a) *perturbation invariance* — for all $\mathbf{p} \in \mathcal{S}^D$, $(\mathbf{p} \circ \mathbf{r})_S = (\mathbf{p} \circ \mathbf{x})_S$;
   (b) *power transformation invariance* — for all $\alpha \in \Re$, $(\alpha \diamond \mathbf{r})_S = (\alpha \diamond \mathbf{x})_S$;
   (c) *subcomposition invariance* — $\mathbf{x}_S = \mathbf{r}_S$.
3. When $\mathbf{x}$ and $\mathbf{x}^*$ have "common" zero values, and the replaced observations $\mathbf{r}$ and $\mathbf{r}^*$ are obtained using identical imputation values $\delta_j = \delta_j^*$, then
   (a) $r_j/r_j^* = x_j/x_j^*$ for all non-zero values x_j, x_j^*, and $d_a(\mathbf{r}, \mathbf{r}^*)$ does not depend on the imputated values;
   (b) $d_a(\mathbf{r}, \mathbf{r}^*)$ is not equal to $d_a(\mathbf{x}_S, \mathbf{x}_S^*)$, but the following equality holds:

$$d_a^2(\mathbf{r}, \mathbf{r}^*) = d_a^2(\mathbf{x}_S, \mathbf{x}_S^*) + \frac{C}{D(D-C)} \left[\sum_{x_j > 0} \log \left(\frac{x_j}{x_j^*} \right) \right]^2 ,$$

   where C is the number of common zeros in $\mathbf{x}$ and $\mathbf{x}^*$.

5 Example

Consider the Glacial data set included in Aitchison (1986). It has 92 samples of pebbles of glacial tills sorted into four categories: red sandstone, gray sandstone, crystalline, and miscellaneous. The components x_1, x_2, x_3, and x_4 represent the corresponding percentages by weight of these four categories. Zeros appear in 41 out of the 92 observations either in component x_3 or in x_4. We assume the zeros to be non-essential zeros, *i. e.* rounded zeros. Before applying a hierarchical clustering algorithm, the zeros have to be replaced. For comparison purposes, we consider the additive replacement approach proposed by Aitchison (1) and the multiplicative replacement (2) proposed in this paper, combined with two different δ values $\delta_1 = 0.001$ and $\delta_2 = 0.0005$. As a consequence, four data sets without zeros are obtained: $\mathbf{R}_{1,1}$ using method (1) and δ_1; $\mathbf{R}_{1,2}$ using the same method but δ_2; $\mathbf{R}_{2,1}$ using method (2) and δ_1; and $\mathbf{R}_{2,2}$ using method (2) and δ_2. The clustering algorithm used has been Ward's method adapted to compositional data (Martín-Fernández et al., 1998), resulting in two distinct groups in all four cases. Comparing the two groups obtained in each case, the following facts can be observed: classifications of $\mathbf{R}_{1,1}$ and $\mathbf{R}_{2,1}$ are extremely coincident, as only one observation is assigned to a different group; classifications of $\mathbf{R}_{2,1}$ and $\mathbf{R}_{2,2}$ are identical; and classification of $\mathbf{R}_{1,2}$ is appreciably different of the rest, as 17 observations are assigned to a different group when compared to $\mathbf{R}_{1,1}$. This indicates that with the multiplicative replacement (2) the matrix of distances is more stable with respect to changes of the imputated values δ_j. But, when the imputated values tend to zero the two replacement sets tend to give us the same results. If we take $\delta = 10^{-8}$ in the two cases and we apply Ward's method, we obtain 4 distinct groups (see Figure 1).

Group G1 corresponds to the observations without zeros, group G2 corresponds to observations with zero only in component x_3, group G3 corresponds to observations with zero only in component x_4, and group G4 corresponds to observations with zeros in both components. These groups could be obtained if initially we assume the zeros as essential zeros rather than rounded zeros.

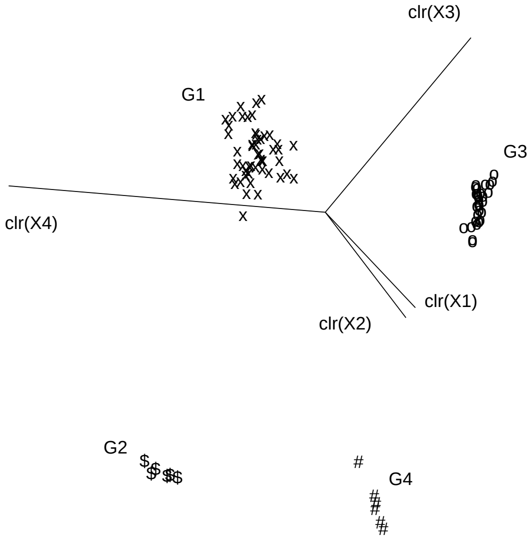

Fig. 1. Biplot of the compositional data set obtained by replacement (2) when $\delta = 10^{-8}$. Labels G1,G2,G3, and G4 represents the four groups.

6 Conclusions

In this paper, a multiplicative zero replacement method for compositional data is defined. This replacement is coherent with the basic operations which provide the simplex with a vector space structure. In particular, the multiplicative approach is "natural" in the sense that it recovers the "true" observation if replacement values are identical to the missing values.

Acknowledgments

This research has been partially supported by the University of Girona through the project UdG98/32, and by the Dirección General de Enseñanza Superior (DGES) of the Spanish Ministry for Education and Culture through the project PB96-0501-C02-01.

References

AITCHISON, J. (1986): *The Statistical Analysis of Compositional Data.* Chapman and Hall, New York (USA), 416 p.

AITCHISON, J., BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J.A., and PAWLOWSKY-GLAHN, V. (2000): Logratio analysis and compositional distance. *Mathematical Geology*, (in press).

KRZANOWSKI, W.J. (1988): *Principles of Multivariate Analysis. A User's Perspective*, Clarendon Press, Oxford (GB), 563 p. (reprinted 1996).

LITTLE, R.J.A and RUBIN, D.B. (1987): *Statistical Analysis with Missing Data.* John Wiley & and Sons, New York (USA), 278 p.

MARTÍN-FERNÁNDEZ, J.A., BARCELÓ-VIDAL, C., and PAWLOWSKY-GLAHN, V. (1998): A Critical Approach to Non-parametric Classification of Compositional Data. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification.* Springer, Heidelberg, pp. 49-56.

SANDFORD, R.F, PIERSON, C.T., and CROVELLI, R.A. (1993): An Objective Replacement Method for Censored Geochemical Data. *Mathematical Geology*, Vol. 25:1, pp 59-80.

TAUBER, F. (1999): Spurious clusters in Granulometric Data Caused by Logratio Transformation. *Mathematical Geology*, Vol. 31:5, pp. 491-504.