A convenient device for replacing rounded zeros in compositional data: *aln* model

Palarea-Albaladejo, J.[1], Daunis-i-Estadella, J.[2] and Martín-Fernández, J.A.[2]

[1] Dept. Informática de Sistemas, Univ. Católica San Antonio, Murcia, Spain.
[2] Dept. Informática y Matemática Aplicada, Univ. de Girona, Girona, Spain.

Poster presentation

1 Introduction

Formally, a composition is a vector $\mathbf{x} = [x_1, \ldots, x_D]$ such that $x_j > 0$, $j = 1, \ldots, D$, subject to the constraint $x_1 + \ldots + x_D = 1$. The sample space of compositions is the unit simplex $\mathcal{S}^D$. Its peculiarities prevent us from applying the standard multivariate statistical techniques designed for real spaces. Log-ratio methodology (Aitchison, 1986) provides the one to one correspondences between the simplex and the real space, opening up the whole of unconstrained real space multivariate data analysis. The results can then be translated back into the compositions of the simplex.
Sometimes in practice some parts take *rounded zero* values or *trace* zeros, making it impossible to use the log-ratio methodology. From a non-parametric point of view, the *multiplicative replacement* (MR) method (Martín-Fernández et al., 2003) replaces the zeros by a small number provided by the analyst. In this work, a computationally feasible parametric method based on a modification of the EM-algorithm is proposed. Its performance is analyzed by Monte Carlo simulation.

2 *aln*: multivariate log-ratio normal model

Aitchison (1986) introduces the additive log-ratio transformation $\mathrm{alr}(\mathbf{x}) = \left[\ln \frac{x_1}{x_D}, \ldots, \frac{x_{D-1}}{x_D}\right] \in R^{D-1}$. Since the alr transformation is asymmetric in the components, one must verify that the applied statistical technique is invariant under permutations of the components. In addition, the alr transformation is not an isometry. To avoid the above difficulties, an isometric log-ratio transformation (ilr) is introduced (Egozcue et al., 2003)

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{y} = [y_1, \ldots, y_{D-1}] \in R^{D-1}, \text{ where } y_i = \frac{1}{\sqrt{i(i+1)}} \ln\left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i}\right),$$

which allows to apply any multivariate technique to the coordinates from an orthonormal basis. In our strategy, the original zeros in the compositional data set $\mathbf{X}$ are transformed in missing data in $\mathbf{Y} = alr(\mathbf{X})$. The main idea is to impute the missing part of $\mathbf{Y}$ and transform back from R^{D-1} to $\mathcal{S}^D$. We select the alr transformation rather than ilr transformation because with the alr transformation the information about the detection limit can be easily incorporated to the alr-transformed data model. Furthermore, the consistency of results is guaranteed (Aitchison, 1986) when inference is based on the likelihood of the additive logistic normal model. Recall that a random composition vector $\mathbf{x} \in S^D$ is distributed according to an additive logistic-normal (aln) model (Aitchison, 1986) when $\mathbf{y} = alr(\mathbf{x})$ is distributed according to a $(D-1)$-dimensional normal model with mean vector μ and covariance matrix Σ.

3 *Modified* EM-algorithm in combination with aln model

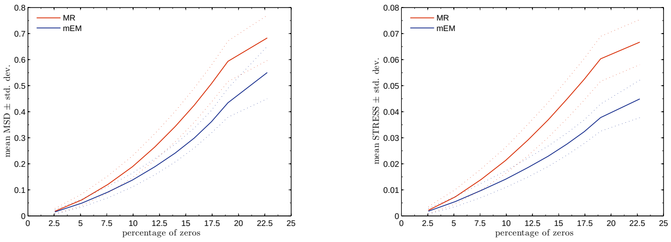
A rounded zero occurs when $x_{ij} < \gamma_j$, where γ_j denotes the detection limit for the component x_j. When this relationship is alr-transformed into the real space, a missing data in $\mathbf{Y}$ is obtained when $y_{ij} < \psi_{ij}$. Note that here $\psi_{ij} = \ln(\gamma_j / x_{iD})$, being x_D a part without zero values. On the t*th* iteration of the *modified* EM-algorithm (mEM) a missing value in the position (i, j) of $\mathbf{Y}$ is imputed (Palarea-Albaladejo et al., 2007a, 2007b) using the equation

$$E[y_j | \mathbf{y}_{-j}, y_j < \psi_j, \theta^{(t)}] = \mathbf{y}_{-j}^T \beta - \sigma_j \frac{\phi\left(\frac{\psi_j - \mathbf{y}_{-j}^T \beta}{\sigma_j}\right)}{\Phi\left(\frac{\psi_j - \mathbf{y}_{-j}^T \beta}{\sigma_j}\right)},$$

where $\theta^{(t)}$ denotes the t*th* estimated parameters vector $\theta = (\mu, \Sigma)$ of the aln model; ϕ and Φ the density and the distribution function, respectively, of the standard normal distribution; σ_j^2 denotes the variance of y_j, and β is the vector of coefficients of the linear regression of y_j on $\mathbf{y}_{-j}$. Note that imputing by this way the method takes into account the information contained in the observed variables as much as the information about the detection limit. The EM algorithm generates a sequence $\{\theta^{(t)}\}$ which converges iteratively (Dempster et al., 1977) to the maximum-likelihood estimate of θ.

4 Simulation-based numerical results

Initially, 1000 data sets of size 300 × 5 are generated from a 5-part random composition. The compositional geometric mean of the random composition $\mathbf{c}$ is given by $g(\mathbf{c}) = [0.027, 0.045, 0.201, 0.605, 0.122]$, and its total



(a) (b)

FIGURE 1. Replacement methods distortion: (a) MSD. (b) STRESS.

variability, totvar($\mathbf{c}$), is equal to 3.996. The value of the geometric mean ensures that the fourth part takes the highest values, and parts 1 and 2 take the smallest values. In addition, the slightly high variability introduced ensures that the simulated data sets not are too similar. These data sets are free of zeros. Following that, small values in the compositions are changed by zero. In this way, a range of 10 realistic detection limits is considered: from 0.25% to 2.5% with increments of 0.25%. Thus, in total, 10 000 data sets of compositional data with rounded zeros have been generated. Next, the data sets are sorted in ascending order according to the proportion of zeros and the MR and mEM strategies to replace them are applied. For multiplicative replacement, the zeros are replaced by the 65% of the corresponding detection limit. MSD and the STRESS

$$\text{MSD} = \frac{\sum_{i=1}^{300} d_a^2(\mathbf{c}_i, \mathbf{r}_i)}{300} \quad \text{and} \quad \text{STRESS} = \frac{\sum_{i<j} (d_a(\mathbf{c}_i, \mathbf{c}_j) - d_a(\mathbf{r}_i, \mathbf{r}_j))^2}{\sum_{i<j} d_a^2(\mathbf{c}_i, \mathbf{c}_j)},$$

evaluate the distortion between the data set $\mathbf{C}$ and the *completed* data set $\mathbf{R}$. By d_a we denote the Aitchison distance between two compositions $\mathbf{x}$ and $\mathbf{x}^*$ defined as the Euclidean distance between the vectors $ilr(\mathbf{x})$ and $ilr(\mathbf{x}^*)$.

Figure 1 shows the patterns followed by the MR and mEM methods in relation to the proportion of zeros in the samples by means of the average of the MSD and STRESS measures (*continuous lines*), ± their respective standard deviations (*dotted lines*), for different intervals of percentages of zeros. When the number of zeros grown the performance of mEM overcome that of MR. Since the MR method replaces all zeros by the same value, it tends to underestimate the variability in the data sets (figure 2). For all samples, the differences between the log-ratio total variabilities for both, the *completed* data set $\mathbf{R}$ and the data set $\mathbf{C}$, is plotted. The mEM method also tends to underestimate the variability since it replaces zeros with an expected value, but this effect is appreciably smaller. With compositions

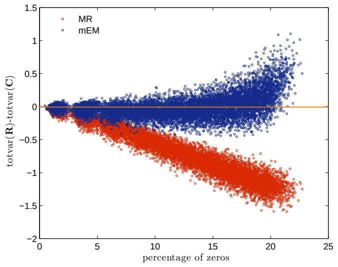


FIGURE 2. Sample variability subestimation.

of higher dimensions the expected result is that the mEM algorithm works better, since the information available to replace zeros by suitable values will increase. The same result will happen if the sample size is enlarged. Therefore, the yield of the mEM algorithm is bound by the size of the data matrices, as is common in all parametric strategies.

Acknowledgments: Work partially financed by the D.G.I. of the Spanish Ministry for Science and Technology (MTM2006-03040).

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* London: Chapman & Hall, 416 pp. Reprinted in 2003 by Blackburn Press.

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

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C. (2003). Isometric log-ratio transformations for compositional data analysis. *Math. Geol.*, **35** , 3, 279-300.

Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets. *Math. Geol.*,**35**, 3, 253-278.

Palarea-Albaladejo, J., Martín-Fernández, J. A., Gómez-García, J. (2007a). A parametric approach for dealing with compositional rounded zeros. *Math. Geol.* (*in press*).

Palarea-Albaladejo, J., Martín-Fernández, J. A., Gómez-García, J. (2007b). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computer and Geosciences*, (*in press*).