An EM-Algorithm Based Method to Deal with Rounded Zeros in Compositional Data under Dirichlet Models

Rafiq Hijazi

Department of Statistics
United Arab Emirates University
P.O. Box 17555, Al-Ain
United Arab Emirates
rhijazi@uaeu.ac.ae

Abstract

Zeros in compositional data are classified into "rounded" zeros and "essential" zeros. The rounded zero corresponds to a small proportion or below detection limit value while the essential zero is an indication of the complete absence of the component in the composition. Several parametric and non-parametric imputation techniques have been proposed to replace rounded zeros and model the essential zeros under logratio model. In this paper, a new method based on EM algorithm is proposed for replacing rounded zeros. The proposed method is illustrated using simulated data.

1 Introduction

Compositional data are non-negative proportions with unit-sum. This type of data arises whenever objects are classified into disjoint categories and the resulting relative frequencies are recorded, or partition a whole measurement into percentage contributions from its various parts. The sample space of compositional data is the simplex $\mathcal{S}^D$ defined as

$$\mathcal{S}^D = \{(x_1, \ldots, x_D) : x_j > 0 \text{ for } j = 1, \ldots, D \text{ and } \sum_{j=1}^{D} x_j = 1\}$$

In compositional data analysis, the presence of zero components represents one of the main obstacles facing the application of both logratio analysis and Dirichlet regression. In a logratio analysis, we cannot take the logarithm of zero when applying the additive logistic transformation. In the Dirichlet model, the presence of zeros makes the probability density function vanish. In this paper, we propose a new technique, based on EM algorithm, for replacing the zeros under Dirichlet model. Section 2 gives an overview of zeros and zero replacement strategies in compositional data. The new EM algorithm based replacement method is described in Section 3. An application to illustrate the use of the proposed technique is presented in Section 4. Finally, concluding remarks are given in Section 5.

2 Zero Replacement in Compositional Data

2.1 Types of Zeros in Compositional Data

Aitchison (1986) classified the zeros in compositional data into "rounded" or trace elements zeros and "essential" or true zeros. The trace zero is an artifact of the measurement process, where observation is recorded as zero when it is below the detection limit (BDL). Such zeros should be treated as missing values and should be replaced. On the other hand, often the observation is recorded as zero as an indication of the complete absence of the component in the composition. Such zeros are called "essential" or true zeros and their pattern of occurrence should be investigated and modeled.

2.2 Common Zero Replacement Methods

As a solution to the rounded zeros problem, Aitchison (1986) suggested the reduction of the number of components in the composition by amalgamation. That is, eliminating the components with zero observations by combining them with some other components. Such approach is not appropriate when

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

the goal is modeling the original compositions or the model includes only three components. However, a more logical approach is to replace the rounded zeros by a small nonzero value that does not seriously distort the covariance structure of the data (Martín-Fernández *et al.* 2003a). The first replacement method, the additive replacement, proposed by Aitchison (1986) is simply replacing the zeros by a small value δ and then normalizing the imputed compositions. Fry et al. (2000) showed that the additive replacement is not subcompositionally coherent and consequently, distorts the covariance structure of the data set. Martín-Fernández et al. (2003) proposed an alternative method using a multiplicative replacement which preserves the ratios of nonzero components. Let $x = (x_1, \ldots, x_D) \in \mathcal{S}^D$ be a composition with rounded zeros. The multiplicative method replaces the composition x containing c zeros with a zero-free composition $r \in \mathcal{S}^D$ according to the following replacement rule

$$r_j = \begin{cases} \delta_j & \text{if } x_j = 0 \\ (1 - c\delta)\, x_j & \text{if } x_j > 0 \end{cases} \tag{1}$$

In addition, Martín-Fernández et al. (2003) emphasized that the best results are obtained when δ is close to 65% of the detection limit. However, since the multiplicative replacement imputes exactly the same value in all the zeros of the compositions, this replacement introduces artificial correlation between components which have zero values in the same composition.

Besides these nonparametric approaches, a parametric approach based on applying a modified EM algorithm on the additive logratio transformation has been proposed (Martín-Fernández et al. (2003), Palarea-Albaladejo et al. (2007) and Palarea-Albaladejo and Martín-Fernández (2008)). However, none of these methods is applicable when the compositional data arise from Dirichlet distribution.

Recently, Hijazi (2009) proposed a new method based on beta regression to replace the rounded zeros when the data arise from Dirichlet distribution. However, the proposed method might rarely replace the rounded zeros by values exceeding the detection limit.

3 EM Algorithm Based Method

The EM (Expectation-Maximization) algorithm is a widely used iterative algorithm for parameter estimation by maximum likelihood method when part of the data is missing or censored data. The algorithm consists of two steps; the E and the M. The E step finds the conditional expectation of the missing data given the observed data and current estimated parameters, and then replaces missing values by estimated values. The M step performs maximum-likelihood estimation of the parameters using estimated parameter values as true values.

In the compositional context, let $X = [x_{ij}]$ be a matrix containing a random sample of n D-component compositions such that $X = (X_{obs}, X_{miss})$ where X_{obs} and X_{miss} are the observed and the missing parts respectively. Assume that X has a Dirichlet distribution with parameters $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_D)$ and a rounded zero occurs when $x_{ij} < \gamma$ where γ is the detection limit. The complete-data log-likelihood for Λ based on a sample X of size n is given by

$$\ell = \log L = n \left\{ \log \Gamma(\lambda) - \sum_{j=1}^{D} \log \Gamma(\lambda_j) \right\} + \sum_{i=1}^{n} \sum_{j=1}^{D} (\lambda_j - 1) \log(x_{ij}) \tag{2}$$

where $\lambda = \sum_j^D \lambda_j$. A useful reparametrization of Dirichlet density is constructed by letting $\mu_j = \lambda_j/\lambda$ for $j = 1, 2, \cdots; D-1$ while λ will be the dispersion parameter of the model.

Since the Dirichlet distribution belongs to the exponential family, the complete data log-likelihood is linear in the non-observed data. Hence, on the t^{th} iteration of the EM algorithm, the computation of the conditional expected complete data likelihood in the E-step reduces to the computation of the conditional expected value of the missing part, $E[X_{miss}|X_{obs}; \theta^{(t)}]$, given a parameter estimate $\theta^{(t)} = \Lambda^{(t)}$.

The E-step in the EM algorithm should be modified to incorporate the detection limit. Hence, on the t^{th} iteration in the E-step, the values of x_{ij} should be replaced according to the following rule:

$$x_{ij}^{(t)} = \begin{cases} x_{ij} & \text{if } x_{ij} \geq \gamma \\ E\left(x_{ij} | \mathbf{x}_{i,-j}, x_{ij} < \gamma, \Lambda^{(t)}\right) & \text{if } x_{ij} < \gamma \end{cases} \tag{3}$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, D$, where $\mathbf{x}_{i,-j}$ denotes the set of observed variables for the i^{th} composition of the data. The conditional expectation in the proposed method should be computed under Dirichlet distribution analogous to the work of Palarea-Albaladejo *et al.* (2007). The expectation in (3) is written as

$$E\left(x_j|\mathbf{x}_{-j}, x_j < \gamma\right) = \frac{1}{P(x_j < \gamma)} \int_0^{\gamma} \frac{\Gamma(\lambda)}{\Gamma(\lambda_j)\Gamma(\lambda - \lambda_j)} x_j^{\lambda_j - 1}(1 - x_j)^{\lambda - \lambda_j - 1} dx_j \tag{4}$$

where λ and λ_j are obtained from the Beta regression of x_j on $\mathbf{x}_{-j}$ (Hijazi, 2009). The latest expression will reduce to

$$E\left(x_j|\mathbf{x}_{-j}, x_j < \gamma\right) = \frac{\lambda_j F_1(\gamma)}{\lambda F_2(\gamma)} \tag{5}$$

where F_1 and F_2 are the cumulative distribution functions of beta random variables with parameters $(\lambda_j + 1, \lambda - \lambda_j)$ and $(\lambda_j, \lambda - \lambda_j)$, respectively.

4 Example: Simulated Data

In this section, we present an application using a simulated data to illustrate the proposed replacement method. The simulated data in Figure 1 consist of 100 compositions randomly generated from $\mathcal{D}(1, 19, 30)$. This paper focuses on the new proposed replacement method, we will consider that the detection limit is 0.5%. This will result in 14 compositions with first component recorded as rounded zero as shown in Figure (1a). The Euclidean distance between the original data and the imputed data using the multiplicative method, beta regression based method and EM algorithm based method are 0.00775, 0.01389 and 0.00768, respectively. This indicates that the new replacement method yielded an imputed data which is slightly closer to the original than the one produced by the multiplicative method. The beta regression based methods appears to perform poorly in this case. The compositions with rounded zeros and the corresponding imputed compositions are shown in Figure (1b). The maximum likelihood estimates of the original data and the imputed data are given in Table (1). It is clear that the EM algorithm based method has resulted in the closest estimates compared to the original data, however, the other two methods have distorted the covariance structure of the original data.

Table 1: Maximum likelihood estimates of original and imputed data

	λ_1	λ_2	λ_3
Original Data	1.183	19.517	32.267
Multiplicative replacement	1.372	21.573	35.701
Beta regression replacement	1.477	22.595	37.409
EM algorithm replacement	1.154	19.278	31.869

5 Conclusion

In this work we have proposed a new replacement method based on EM algorithm under Dirichlet model. The proposed method was compared with the multiplicative replacement and the beta regression based methods through an illustrative example. The new method outperforms the multiplicative replacement and beta regression based methods. This method gives positive imputed value and takes into account the detection limit of the part. A Monte Carlo simulation study should be conducted to evaluate the performance of the new technique with varying covariance structures and percentage of rounded zeros.

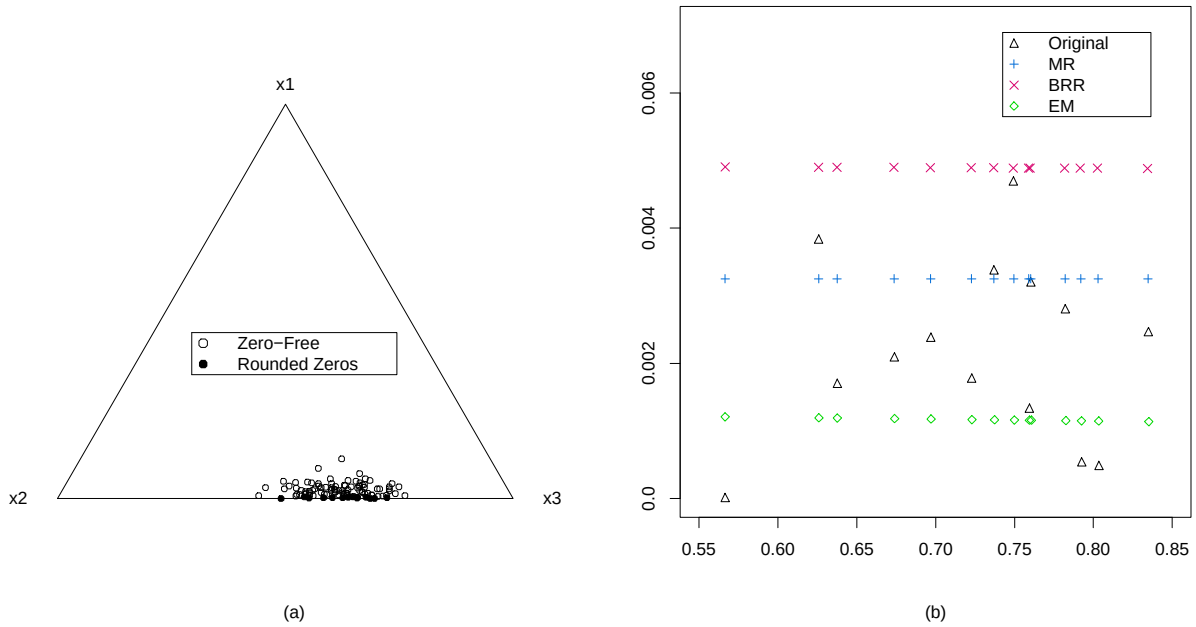

Figure 1: (a) The ternary diagram of the simulated data (b) The original compositions before rounding, imputed compositions with MR (multiplicative method), BRR (beta regression based method) and imputed compositions with EM (EM algorithm based method)

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Fry, J. M., Fry, T. R. L., and McLaren, K. R. (2000). Compositional data analysis and zeros in micro data. *Applied Economics 32*(8), 953–959.

Hijazi R. (2009). Dealing with Rounded Zeros in Compositional Data under Dirichlet Models. *Proceedings of the 10th Islamic Countries Conference on Statistical Sciences 2009 (ICCS-X).*

Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V. (2003). Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology 35*(3), 253–278.

Martín-Fernández, J.A., Palarea-Albaladejo J, Gómez-García, J. (2003). Markov Chain Monte Carlo Method Applied to Rounding Zeros of Compositional Data: First Approach. *Proceedings of CODAWORK'05, The 2nd Compositional Data Analysis Workshop.*

Palarea-Albaladejo, J., Martín-Fernández, J. (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences 34*, 902–917.

Palarea-Albaladejo, J., Martín-Fernández, J., Gómez-García, J. (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology 39*, 625–645.