Hacettepe Journal of Mathematics and Statistics
Volume 48 (1) (2019), 242 – 254

A kernel density approach for replacing rounded zeros in compositional data sets

Jiajia Chen*†, Xiaoqin Zhang‡ and Shengjia Li§

Abstract

The logratio methodology widely used in compositional data analysis is not applicable when some components have rounded zeros. There are many univariate and multivariate methods that have been used to deal with rounded zeros. However, both of them have restrictions: the univariate methods replaced the rounded zeros only using the information of the corresponding component; the multivariate methods need to assume the distribution of transformed data. When the form of the distribution function is unknown, a multivariate nonparametric replacement approach is proposed in this paper. The proposed method uses conditional expected value based on isometric logratio coordinates to replace rounded zeros, in which the conditional density is estimated through multivariate Gauss kernel function. The permutation invariance and invariance under change of orthonormal basis are also presented. Simulation studies show that the proposed method has better performance than previous methods as the percentage of rounded zeros increases. The proposed method is also applied on the moss data from the Kola project.

Keywords: Compositional data, Isometric logratio coordinates, Rounded zeros, Gauss kernel function, Detection limit.

2000 AMS Classification: 62-07, 62G07, 62H11.

Received : 10.07.2017 *Accepted* : 28.06.2018 *Doi* : 10.15672/HJMS.2018.605

*School of Statistics, Shanxi University of Finance and Economics, Taiyuan 030006, Shanxi, P.R. China Email: chenjiajia0401@163.com

†Corresponding Author.

‡School of Mathematical Sciences, Shanxi University, Taiyuan 030006, Shanxi, P.R. China Email: zhangxiaoqin@sxu.edu.cn

§School of Mathematical Sciences, Shanxi University, Taiyuan 030006, Shanxi, P.R. China Email: shjiali@sxu.edu.cn

1. Introduction

Compositional data, or compositions, are vectors in which all components are positive real numbers and carry only relative information [1]. These vectors can be represented as proportions using closure operation, that is, they multiplied by the appropriate scaling factors. Two vectors are compositional equivalent if they can be expressed in the same proportion, thus compositions can be viewed as equivalence classes, in which all vectors convey the same compositional information [18]. This type of data often occurs in geosciences, biosciences, economics and many other disciplines [1, 16, 18].

Compositional data provide information only about the relative magnitudes of the components, the logratio methodology plays a key role in compositional data analysis. Three logratio transformations including additive logratio (alr) transformation [1], centered logratio (clr) transformation [1] and isometric logratio (ilr) transformation [6] were proposed. The relationship between alr transformation and clr transformation is well known [1], and ilr transformation can be represented by means of alr transformation or clr transformation [6]. Because the alr transformation is non-isometric, and the clr transformation results in singular covariance matrix, the ilr transformation which can avoid the above drawbacks is suggested. The logratio transformations transform compositional data to coordinates in real space. However, zeros may exist in some components, thus the logratio transformations fail.

There are three kinds of zeros in compositional data set: rounded zeros, count zeros and essential zeros [9]. In this paper, we are interested in the rounded zero which is not true zero and results from the existence of value below a threshold. When the threshold is rounding-off error, the component is present in a very small quantity and rounded to zero; when the threshold is detection limit, the value below the detection limit cannot be observed and is commonly reported as zero. There are many classic methods in rounded zeros problem. Aitchison proposed the additive replacement strategy [1], but the ratios of components having no rounded zeros are not preserved, later the multiplicative replacement strategy [8] was proposed. Instead of replacing rounded zeros in a component by a fixed value, the multiplicative lognormal replacement method [13] allowing for random imputation was suggested. The multivariate method is the modified EM algorithm [15, 12], which assumed that the alr coordinates follow multivariate normal distribution. Later the robust modified EM algorithm working on ilr coordinates [10] was introduced. In addition, there are other algorithms, for example, the multiplicative Kaplan-Meier method [14] was proposed, which is a univariate method. The implementations of all these methods discussed above are available in the R package zCompositions [14].

The previous univariate methods replace rounded zeros based on the data of the corresponding component and perform poorly when the proportion of rounded zeros is high. The multivariate methods for rounded zeros usually rely on the underlying assumption of multivariate normality in the space of coordinates. Furthermore, the modified EM algorithm based on alr coordinates requires that at least one component has no rounded zeros. To avoid these disadvantages, a new multivariate nonparametric replacement method based on multivariate Gauss kernel density estimation is proposed in this paper. To illustrate the performance of proposed method compared with the existing methods, this method is applied to both simulation and example analysis.

The rest of this paper is organized as follows. Some basic concepts about compositional data are reviewed in Section 2. In Section 3, the proposed approach is presented. Simulation study and real example are given in Section 4 to verify the effectiveness and usefulness of proposed method. Section 5 concludes this paper.

2. Preliminaries

Let $\mathbf{x} = [x_1, x_2, \cdots, x_D]$ be a row vector denoting a D-part composition represented with constant sum k, its sample space is the simplex $\mathcal{S}^D$ [1] defined as

$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, x_2, \cdots, x_D] \middle| x_i > 0, i = 1, 2, \cdots, D; \sum_{i=1}^{D} x_i = k \right\},$$

where the constant k is an arbitrary positive real number and is usually 1 or 100 depending on the units of measurement. The simplex is a Euclidean vector space structure [1, 17, 3] when defining inner product with its related norm and Aitchison distance [2]. The distance between two compositions $\mathbf{x}$ and $\mathbf{y} \in \mathcal{S}^D$ is

$$d_a(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{D} \left(\ln \frac{x_i}{g_m(\mathbf{x})} - \ln \frac{y_i}{g_m(\mathbf{x})} \right)^2 \right)^{1/2},$$

where $d_a(\cdot, \cdot)$ stands for the Aitchison distance in $\mathcal{S}^D$, and $g_m(\mathbf{x})$ denotes the geometric mean of the parts of $\mathbf{x}$.

The ilr transformation [6] assigns coordinates with respect to the given orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{D-1}\}$ of the simplex $\mathcal{S}^D$. An orthonormal basis can be obtained through sequential binary partition of parts of a composition [5]. Following the reference [5], we can construct a $(D-1) \times D$ matrix $\mathbf{\Psi}$ in which rows are

$$\psi_i = \sqrt{\frac{D-i}{D-i+1}} \left[0, \cdots, 0, 1, \underbrace{-\frac{1}{D-i}, \cdots, -\frac{1}{D-i}}_{D-i} \right], \quad i = 1, 2, \cdots, D-1, \tag{2.1}$$

respectively. An orthonormal basis can be obtained through $\mathbf{e}_i = \mathcal{C}(\exp\ \psi_i)$ $(i = 1, 2, \cdots, D-1)$, where $\mathcal{C}$ is the closure operation. Thus the composition $\mathbf{x} \in \mathcal{S}^D$ is transformed to ilr coordinates $\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = [z_1, z_2, \cdots, z_{D-1}] \in \mathbb{R}^{D-1}$, where

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, 2, \cdots, D-1. \tag{2.2}$$

The ilr coordinates guarantee the invariance of distance, that is, $d_a(\mathbf{x}, \mathbf{y}) = d(\mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y}))$, where $d(\cdot, \cdot)$ is the Euclidean distance in real space. The inverse mapping of any real-valued vector $\mathbf{z} \in \mathbb{R}^{D-1}$ to the original composition $\mathbf{x}$ is then given by

$$\begin{cases} x_1 = \exp\left\{ \sqrt{\frac{D-1}{D}} z_1 \right\}, \\ x_i = \exp\left\{ -\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \sqrt{\frac{D-i}{D-i+1}} z_i \right\}, \quad i = 2, \cdots, D-1, \\ x_D = \exp\left\{ -\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j \right\}. \end{cases} \tag{2.3}$$

The compositions can be viewed as equivalence classes, therefore the obtained composition $\mathbf{x}$ can be represented as constant sum vectors.

For any random composition $\mathbf{x} = [x_1, x_2, \cdots, x_D]$, the measure of dispersion is the variation matrix [1] defined as

$$\mathbf{T} = [t_{ij}]_{D\times D}, \quad t_{ij} = \mathrm{Var}\left(\ln \frac{x_i}{x_j}\right), \tag{2.4}$$

where the element in variation matrix is the logratio variance for any two parts i and j of a D-part composition $\mathbf{x}$.

3. Kernel density replacement approach

Consider a random composition $\mathbf{x} = [x_1, x_2, \cdots, x_D]$, the sample data set is $\mathbf{X}$ with n compositions and D-part, that is

$$\mathbf{X} = [x_{ij}]_{n\times D} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nD} \end{pmatrix}.$$

Suppose that the compositional data set $\mathbf{X}$ has rounded zeros, the corresponding threshold matrix is denoted as $\mathbf{E} = [e_{ij}]_{n\times D}$, where e_{ij} is the threshold of x_{ij}. Let $R_j \subset \{1, 2, \cdots, n\}$ be the row indices referring to the rounded zeros of the jth component ($j \in \{1, 2, \cdots, D\}$), then $O_j = \{1, 2, \cdots, n\}\backslash R_j$ refers to the remaining row indices of the jth component, that is, $R_j = \{i : i \in \{1, 2, \cdots, n\}, x_{ij} \le e_{ij}\}$, $O_j = \{i : i \in \{1, 2, \cdots, n\}, x_{ij} > e_{ij}\}$.

Firstly, we initialize the rounded zeros by multiplicative replacement strategy in which the rounded zero is equal to 65% of the threshold [8], thus $\mathbf{X}$ denotes the replaced data set. Denote the ilr coordinates in Equation (2.2) of random composition $\mathbf{x}$ as $\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = [z_1, z_2, \cdots, z_{D-1}] = [z_1, \mathbf{z}_{-1}]$, where $\mathbf{z}_{-1}$ refers to the remaining components of $\mathbf{z}$ except for the first component. Then initialized data set $\mathbf{X}$ is transformed to real data set $\mathbf{Z} = [z_{ij}]_{n\times(D-1)}$, where each row in $\mathbf{Z}$ is the ilr coordinates of the corresponding composition in $\mathbf{X}$. For the element e_{i1} in threshold set $\mathbf{E}$, the ilr transformation of rounded zero $x_{i1} < e_{i1}$ can result in the the unknown value z_{i1} less than ψ_{i1}, where

$$\psi_{i1} = \sqrt{\frac{D-1}{D}} \ln \frac{e_{i1}}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}}}.$$

In the proposed approach, the unknown data z_{i1} ($i \in R_1$) is imputed by conditional expected value

$$E(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1}, z_1 < \psi_{i1}) = \frac{\int_{-\infty}^{\psi_{i1}} z_1 f(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1})\mathrm{d}z_1}{\int_{-\infty}^{\psi_{i1}} f(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1})\mathrm{d}z_1}, \tag{3.1}$$

where $\mathbf{z}_{i,-1}$ is the ith row of $\mathbf{Z}$ except for the first column, the conditional density function $f(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1})$ can be calculated as follows

$$f(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1}) = \frac{f(z_1, \mathbf{z}_{-1} = \mathbf{z}_{i,-1})}{f(\mathbf{z}_{-1} = \mathbf{z}_{i,-1})} = \frac{f(z_1, \mathbf{z}_{-1} = \mathbf{z}_{i,-1})}{\int_{-\infty}^{+\infty} f(z_1, \mathbf{z}_{-1} = \mathbf{z}_{i,-1})\mathrm{d}z_1}. \tag{3.2}$$

Regardless the distribution of multivariate random variable $\mathbf{z}$, the density function $f(z_1, \mathbf{z}_{-1} = \mathbf{z}_{i,-1})$ can be estimated by multivariate Gauss kernel density [4]. In this

paper, the same bandwidth h is applied to different coordinate direction, thus

$$\begin{aligned}
&\widehat{f}(z_1, \mathbf{z}_{-1} = \mathbf{z}_{i,-1}) \\
= \quad & \frac{1}{n(\sqrt{2\pi}h)^{D-1}} \sum_{k=1}^{n} \exp\left\{ -\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2 \right. \\
& \left. -\frac{1}{2}\left(\frac{\mathbf{z}_{i,-1} - \mathbf{z}_{k,-1}}{h}\right)\left(\frac{\mathbf{z}_{i,-1} - \mathbf{z}_{k,-1}}{h}\right)^T \right\} \\
= \quad & \frac{1}{n(\sqrt{2\pi}h)^{D-1}} \sum_{k=1}^{n} \exp\left\{ -\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2 - \frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1}) \right\}. \qquad (3.3)
\end{aligned}$$

The bandwidth h is given by $h = \sigma\left(\frac{4}{n(D+1)}\right)^{\frac{1}{D+3}}$ [20], where $\sigma^2 = \frac{1}{D-1}\sum_{j=1}^{D-1} \mathrm{Var}(z_j) = \frac{1}{D-1}\mathrm{tr}(\mathrm{Var}(\mathbf{z}))$, tr represents the trace of matrix $\mathrm{Var}(\mathbf{z})$.

It follows from Equation (3.2) and Equation (3.3) that

$$\begin{aligned}
&\widehat{f}(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1}) \\
= \quad & \frac{\sum_{k=1}^{n} \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \exp\left\{-\frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})\right\}}{\sum_{k=1}^{n} \exp\left\{-\frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})\right\} \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \mathrm{d}z_1} \\
= \quad & \frac{\sum_{k=1}^{n} \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \exp\left\{-\frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})\right\}}{\sum_{k=1}^{n} \sqrt{2\pi}h \exp\left\{-\frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})\right\}}, \qquad (3.4)
\end{aligned}$$

By conditional density function in Equation (3.4), Equation (3.1) can be expressed as

$$\frac{\sum_{k=1}^{n} \exp\left\{-\frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})\right\} \int_{-\infty}^{\psi_{i1}} z_1 \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \mathrm{d}z_1}{\sum_{k=1}^{n} \exp\left\{-\frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})\right\} \int_{-\infty}^{\psi_{i1}} \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \mathrm{d}z_1}. \qquad (3.5)$$

Since

$$\int_{-\infty}^{\psi_{i1}} \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \mathrm{d}z_1 = \sqrt{2\pi}h\Phi\left(\frac{\psi_{i1} - z_{k1}}{h}\right)$$

and

$$\begin{aligned}
& \int_{-\infty}^{\psi_{i1}} z_1 \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \mathrm{d}z_1 \\
= \quad & \int_{-\infty}^{\psi_{i1}} (z_1 - z_{k1}) \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \mathrm{d}z_1 + z_{k1} \int_{-\infty}^{\psi_{i1}} \exp\left\{-\frac{1}{2}\left(\frac{z_1 - z_{k1}}{h}\right)^2\right\} \mathrm{d}z_1 \\
= \quad & \sqrt{2\pi}h\left(-h\phi\left(\frac{\psi_{i1} - z_{k1}}{h}\right) + z_{k1}\Phi\left(\frac{\psi_{i1} - z_{k1}}{h}\right)\right),
\end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function of the standard normal distribution, respectively. Thus Equation (3.5) can be simplified as

$$\begin{aligned}
& E(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1}, z_1 < \psi_{i1}) = \\
& \frac{\sum_{k=1}^{n} \left(-h\phi\left(\frac{\psi_{i1} - z_{k1}}{h}\right) + z_{k1}\Phi\left(\frac{\psi_{i1} - z_{k1}}{h}\right)\right) \exp\left\{-\frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})\right\}}{\sum_{k=1}^{n} \Phi\left(\frac{\psi_{i1} - z_{k1}}{h}\right) \exp\left\{-\frac{1}{2h^2} d^2(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})\right\}}. \qquad (3.6)
\end{aligned}$$

Hence, the unknown data z_{i1} is imputed by Equation (3.6). For the ilr coordinates in Equation (2.2), since $d(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1}) = d_a(\mathbf{x}_{i,-1}, \mathbf{x}_{k,-1})$, the imputed value z_{i1} is related with the Aitchison distance between subcompositions $\mathbf{x}_{i,-1}$ and $\mathbf{x}_{k,-1}$, where $\mathbf{x}_{i,-1}$ and $\mathbf{x}_{k,-1}$ denote the remaining components of compositions $\mathbf{x}_i$ and $\mathbf{x}_k$ except for the first component, respectively.

3.1. Property. *The imputed value* $E(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1}, z_1 < \psi_{i1})$ *in Equation* (3.6) *has the following properties:*

(1) *It is below the threshold, that is,* $E(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1}, z_1 < \psi_{i1}) < \psi_{i1}$.
(2) *It is unchanged when the remaining components of* $\mathbf{x}$ *except for the first component are arbitrarily permuted.*
(3) *It is invariant under change of orthonormal basis* $\{\mathbf{e}_2, \mathbf{e}_3, \cdots, \mathbf{e}_{D-1}\}$.

Property 3.1 is quite obvious. It follows from $\mathbf{z} = \mathbf{x}\mathbf{\Psi}^T$ that $\mathrm{tr}(\mathrm{Var}(\mathbf{z})) = \mathrm{tr}(\mathrm{Var}(\mathbf{x}\mathbf{\Psi}^T)) = \mathrm{tr}(\mathbf{\Psi}\mathrm{Var}(\mathbf{x})\mathbf{\Psi}^T) = \mathrm{tr}(\mathrm{Var}(\mathbf{x})\mathbf{\Psi}^T\mathbf{\Psi}) = \mathrm{tr}(\mathrm{Var}(\mathbf{x})\mathbf{G}_D)$ [18], where $\mathbf{G}_D = \mathbf{I}_D - \frac{1}{D}\mathbf{J}_D$, $\mathbf{I}_D$ is a identity matrix, $\mathbf{J}_D$ is a matrix of units. Therefore all the underlaying elements ($d(\mathbf{z}_{i,-1}, \mathbf{z}_{k,-1})$, h, z_{k1} and ψ_{i1}) are invariant by permutation and change of basis, thus the imputed value in Equation (3.6) is unchanged.

Property 3.1 (2) and (3) point out that $E(z_1|\mathbf{z}_{-1} = \mathbf{z}_{i,-1}, \mathbf{z}_1 < \psi_{i1})$ satisfies permutation invariance and invariance under change of orthonormal basis, but $E(z_l|\mathbf{z}_{-l} = \mathbf{z}_{i,-l}, z_l < \psi_{il})$ $(l = 2, \cdots, D-1)$ may not satisfy these two properties, for example, z_{kl} may changed when the remaining components of $\mathbf{x}$ except for the lth component are arbitrarily permuted. To replace the rounded zeros in the lth component of $\mathbf{x}$, we define the permuted composition $\mathbf{x}^{(l)} = [x_1^{(l)}, x_2^{(l)}, \cdots, x_l^{(l)}, x_{l+1}^{(l)}, \cdots, x_D^{(l)}] = [x_l, x_1, \cdots, x_{l-1}, x_{l+1}, \cdots, x_D]$. The ilr coordinates are denoted as $\mathbf{z}^{(l)} = \mathrm{ilr}(\mathbf{x}^{(l)}) = [z_1^{(l)}, z_2^{(l)}, \cdots, z_{D-1}^{(l)}] = [z_1^{(l)}, \mathbf{z}_{-1}^{(l)}]$, the corresponding ilr data set is $\mathbf{Z}^{(l)} = [z_{ij}^{(l)}]_{n\times(D-1)}$. According to Equation (3.6), the unknown data $z_{i1}^{(l)}$ $(i \in R_l)$ resulting from the rounded zero in the ith row and the lth component of $\mathbf{X}$ can be imputed by

$$E(z_1^{(l)}|\mathbf{z}_{-1}^{(l)} = \mathbf{z}_{i,-1}^{(l)}, z_1^{(l)} < \psi_{i1}^{(l)}) = \frac{\sum_{k=1}^{n}\left(-h\phi\left(\frac{\psi_{i1}^{(l)} - z_{k1}^{(l)}}{h}\right) + z_{k1}^{(l)}\Phi\left(\frac{\psi_{i1}^{(l)} - z_{k1}^{(l)}}{h}\right)\right)\exp\left\{-\frac{1}{2h^2}d^2(\mathbf{z}_{i,-1}^{(l)}, \mathbf{z}_{k,-1}^{(l)})\right\}}{\sum_{k=1}^{n}\Phi\left(\frac{\psi_{i1}^{(l)} - z_{k1}^{(l)}}{h}\right)\exp\left\{-\frac{1}{2h^2}d^2(\mathbf{z}_{i,-1}^{(l)}, \mathbf{z}_{k,-1}^{(l)})\right\}}, \tag{3.7}$$

where $\psi_{i1}^{(l)} = \sqrt{\frac{D-1}{D}}\ln\frac{e_{il}}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}^{(l)}}}$.

The specific steps of the proposed method, similar to the modified EM algorithm based on ilr coordinates [10], are as follows:

Step 1: Sort the parts of compositional data set according to the number of rounded zeros of each part. The ilr coordinates in Equation (2.2) is used in the proposed method, the first component is only included in the first ilr coordinate. In order to reduce the error, the component with more zeros should be put in the first column. Without loss of generality, assume that the parts are already sorted, i.e. $|R_1| \geq |R_2| \geq \cdots \geq |R_D|$, where $|R_j|$ denotes the number of elements of R_j $(j = 1, 2, \cdots, D)$.
Step 2: Initialize the rounded zeros by multiplicative replacement strategy.
Step 3: Set $l = 1$.
Step 4: Replace the unknown data $z_{i1}^{(l)}$ $(i \in R_l)$ using Equation (3.7).
Step 5: Inverse the every row of updated data set using Equation (2.3).

Step 6: Carry out Steps 4-5 for each $l = 2, 3, \cdots |C|$, where $C = \{j : j \in \{1, 2, \cdots, D\}, |R_j| \neq 0\}$ is the index set of parts containing at least one rounded zero.

Step 7: Repeat Steps 3-6 until the Euclidean distance between the variation matrix of compositional data set from the present and the previous iteration is smaller than a certain boundary.

Step 8: Sort the parts of replaced compositional data set in the original order.

If the data set $\mathbf{X} = [x_{ij}]_{n\times D}$ is closed to a constant, then the replaced data set is $\widehat{\mathbf{X}} = [\hat{x}_{ij}]_{n\times D}$ obtained from the above algorithm, otherwise, we should rescale the replaced value $\hat{x}_{ij}$ using the expression [14]

$$\hat{x}^*_{ij} = \hat{x}_{ij}\frac{x_{ik}}{\hat{x}_{ik}}, \quad j \in C,\ i \in R_j, \tag{3.8}$$

where $\hat{x}^*_{ij}$ is the rescaled value, x_{ik} is the originally observed element in the ith row and kth column of compositional data set $\mathbf{X}$, $\hat{x}_{ik}$ is the corresponding replaced value in $\widehat{\mathbf{X}}$.

4. Simulation and Example

In this section we present simulation study and real example in order to illustrate the good performance of proposed method (multK), which is compared with the multiplicative replacement strategy (multR), the multiplicative Kaplan-Meier method (multKM), the multiplicative lognormal replacement method (multLN), the modified EM algorithm working on alr coordinates (alrEM) and the robust modified EM algorithm working on ilr coordinates (ilrEM). Given the original compositional data set $\mathbf{X}$ which has no rounded zeros, we set the value below the threshold as zero, the replaced compositional data set is denoted as $\mathbf{X}^*$. We consider two measures of distortion, standardized residual sum of squares (STRESS) [8] and relative difference in variation matrix (RDVM) [13]. Denote the variation matrix in Equation (2.4) of original data set $\mathbf{X}$ and imputed data set $\mathbf{X}^*$ as $\mathbf{T} = [t_{ij}]_{D\times D}$ and $\mathbf{T}^* = [t^*_{ij}]_{D\times D}$, the two measures STRESS and RDVM are defined as

$$\text{STRESS} = \frac{\sum_{i<j}(d_a(\mathbf{x}_i, \mathbf{x}_j) - d_a(\mathbf{x}^*_i, \mathbf{x}^*_j))^2}{\sum_{i<j} d^2_a(\mathbf{x}_i, \mathbf{x}_j)},$$

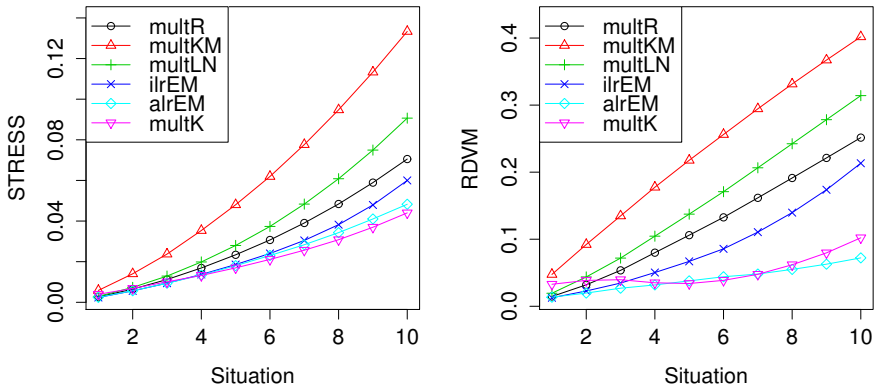
and

$$\text{RDVM} = \frac{1}{2|C|D - |C|^2}\sum_{i,j\in C}\frac{|t^*_{ij} - t_{ij}|}{t_{ij}},$$

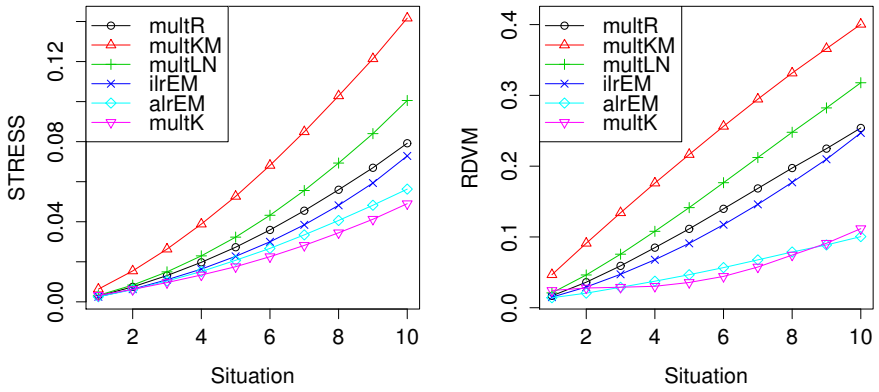
respectively, where $\mathbf{x}_i$ is the ith row of data set $\mathbf{X}$. The two measures STRESS and RDVM represent the distance difference and variation difference, respectively.

4.1. Simulation Study. In this subsection, several simulation studies were conducted. We first simulated real data set with sample size 300 from multivariate normal distribution $N_4(\mu, \mathbf{\Sigma})$, then the compositional data set $\mathbf{X}$ can be obtained through ilr-inverse transformation in Equation (2.3). Suppose that the rounded zero is resulting from value below the detection limit, and the detection limits of same part-different compositions are the same, so the detection limit set is denoted as a vector, that is, $\mathbf{E} = [e_1, e_2, \cdots, e_5]$, where e_j $(j = 1, 2, \cdots 5)$ is the α_j quantile of the jth component in $\mathbf{X}$.

We set mean $\mu = [-2, -1.5, -1, -0.3]$ and covariance matrix $\mathbf{\Sigma} = [\rho^{|i-j|}]_{4\times 4}$. To describe different levels of correlations among the components, take $\rho = 0.3, 0.5, 0.7$ and 0.9. Ten situations of detection limit set are conducted, where α_1, α_2, α_3, α_4 range from 0.05 to 0.5 by 0.05, 0.04 to 0.4 by 0.04, 0.03 to 0.3 by 0.03, 0.02 to 0.2 by 0.02, respectively,



(a) $\rho = 0.3$.

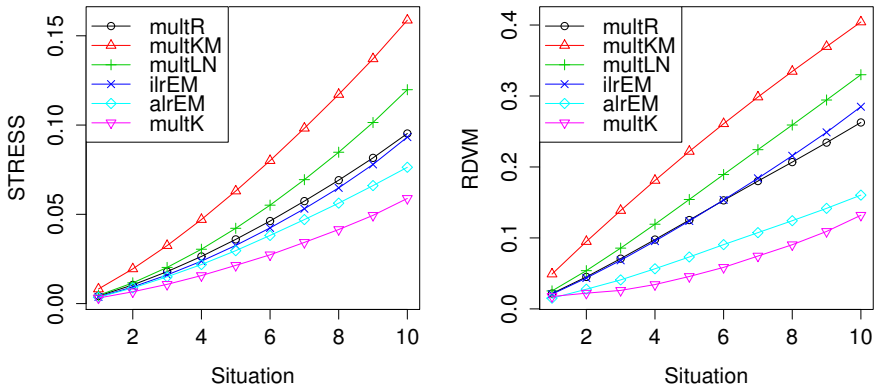


(b) $\rho = 0.5$.

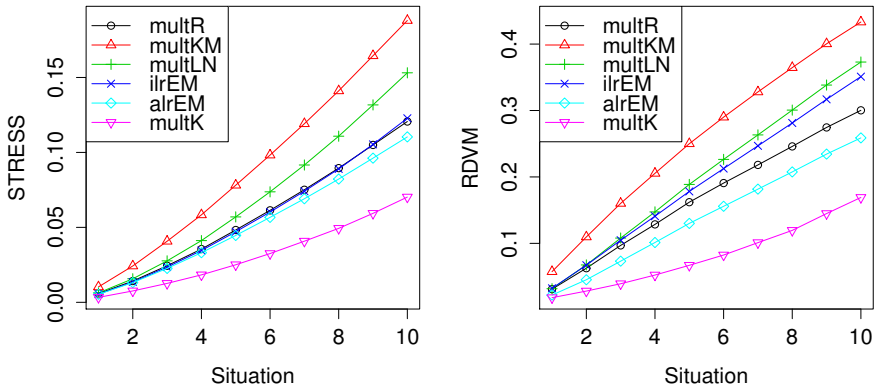
Figure 1. Two measures of distortion STRESS and RDVM for six methods (multR, multKM, multLN, ilrEM, alrEM, multK) under ten situations of detection limit set when $\rho = 0.3$ (a) and $\rho = 0.5$ (b).

and $\alpha_5 = 0$. Set each data in the jth component smaller than e_j $(j = 1, 2, 3, 4)$ to a zero value, then the percentage of rounded zeros in the first four components approximately range from 5% to 50% by 5%, 4% to 40% by 4%, 3% to 30% by 3%, 2% to 20% by 2%, respectively, and the last component has no rounded zeros, therefore the corresponding percentage of rounded zeros in compositional data set approximately ranges from 2.8% to 28% by 2.8%.

We run 100 Monte Carlo simulations for each setting described above. The performance comparisons among previous methods and proposed method with varying percentage of rounded zeros corresponding to situations are showed in Figure 1 and Figure



(a) $\rho = 0.7$.



(b) $\rho = 0.9$.

Figure 2. Two measures of distortion STRESS and RDVM for six methods (multR, multKM, multLN, ilrEM, alrEM, multK) under ten situations of detection limit set when $\rho = 0.7$ (a) and $\rho = 0.9$ (b).

2. The values in Figure 1 and Figure 2 are the average STRESS or RDVM of 100 simulations. Figure 1(a) and Figure 1(b) depict the trends in two performance measures under ten situations of detection limit set when $\rho = 0.3$ and 0.5. It can be seen from Figure 1(a) and Figure 1(b) that the ilrEM and alrEM have smaller STRESS and RDVM than those of multR, however, the STRESS and RDVM of multKM and multLN are greater than those of multR. Moreover, when the percentage of rounded zeros increases, the STRESS value of multK is lower than those of previous methods. The multK method performs worse than previous methods in the measure RDVM when $\rho = 0.3$, whereas it performs better under some situations when $\rho = 0.5$. Figure 2 shows the trends in two measures

among different methods when $\rho = 0.7$ and 0.9. From Figure 2(a) and Figure 2(b), we see that the multK method outperforms the other methods in two measures STRESS and RDVM. The STRESS value of ilrEM is very close to that of multR, while ilrEM performs worse than multR in measure RDVM. To sum up, when the percentage of rounded zeros increases, the proposed method has better performance than other methods in the two measures STRESS and RDVM.

4.2. Real example. The proposed method discussed in the previous section will be applied to the moss data from the Kola project available in the R package StatDA [7] and compared with the previous methods (multR, multKM, multLN, ilrEM and alrEM). The moss data set consists of more than 50 chemical elements and 594 observations. We focus on the 7-part subcomposition [Al, Ca, Fe, K, Mg, Na, Si] denoted as compositional data set $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_7]$ with constant sum 100%, which has no rounded zeros. Similar to the simulation analysis, we give the detection limit set, the value below detection limit is set as zero. The aim of this study is to replace rounded zeros using different methods.

Suppose that the components $\mathbf{u}_1, \mathbf{u}_3, \mathbf{u}_6$ and $\mathbf{u}_7$ have rounded zeros. Eight situations of detection limit set are given in Table 1 in which e_j $(j = 1, 3, 6, 7)$ is the detection limit of the jth component. Table 1 also gives the percentages of rounded zeros of components $\mathbf{u}_1, \mathbf{u}_3, \mathbf{u}_6, \mathbf{u}_7$ and the total percentage of rounded zeros of compositional data set $\mathbf{U}$. Table 2 gives the computed results of STRESS and RDVM for six methods (multR, multKM, multLN, ilrEM, alrEM, multK) under eight situations. According to Table 2, we can find that the proposed method has smaller STRESS value than those of other methods except the first two situations, and the RDVM value of proposed method for each situation is always smaller than other methods. In addition, multR performs better than ilrEM and alrEM as the percentage of rounded zeros increases, of which alrEM has larger STRESS and RDVM than ilrEM. This is because that the ilrEM and alrEM all assume the distribution of compositional data set. In fact, compositional data set $\mathbf{U}$ departures from normal distribution on the simplex [11], which is tested using the energy test [19] or the test based on SVD including the marginal univariate tests, the bivariate tests and radius tests [21]. Because the ilrEM is a robust method, which performs better than alrEM. These results suggest that the proposed method is superior to the others in the case of moss data set.

Table 1. Eight situations of detection limit set for compositional data set **U**. The value in parentheses is the percentage of rounded zeros of the corresponding component. The last column ZR represents the total percentage of rounded zeros of the corresponding situation (Unit: %).

situation	e_1	e_3	e_6	e_7	ZR
1	1.39(14.14)	1.41(13.97)	0.41(14.65)	1.41(14.31)	8.15
2	1.51(18.86)	1.56(18.69)	0.46(19.53)	1.55(19.19)	10.89
3	1.63(23.57)	1.72(23.23)	0.51(24.41)	1.68(23.91)	13.59
4	1.76(28.28)	1.85(27.95)	0.56(29.29)	1.76(28.62)	16.31
5	1.84(33.00)	1.96(32.49)	0.60(34.18)	1.86(33.50)	19.02
6	1.93(37.71)	2.04(37.21)	0.66(39.06)	1.98(38.22)	21.74
7	2.01(42.42)	2.22(41.75)	0.72(43.94)	2.05(42.93)	24.43
8	2.12(47.14)	2.38(46.46)	0.78(48.82)	2.13(47.81)	27.18

Table 2. Two evaluation indexes STRESS and RDVM of methods (multR, multKM, multLN, ilrEM, alrEM, multK) for compositional data set **U** under eight situations of detection limit set.

	situation	multR	multKM	multLN	ilrEM	alrEM	multK
STRESS	1	0.0179	0.0317	0.0166	0.0148	0.0158	0.0159
	2	0.0216	0.0426	0.0213	0.0182	0.0212	0.0189
	3	0.0244	0.0567	0.0275	0.0222	0.0260	0.0218
	4	0.0283	0.0706	0.0348	0.0265	0.0336	0.0257
	5	0.0328	0.0833	0.0422	0.0312	0.0444	0.0302
	6	0.0372	0.0994	0.0518	0.0372	0.0592	0.0358
	7	0.0425	0.1185	0.0638	0.0468	0.0737	0.0421
	8	0.0494	0.1382	0.0774	0.0613	0.1042	0.0493
RDVM	1	0.0623	0.1626	0.0645	0.0478	0.0465	0.0389
	2	0.0671	0.2033	0.0864	0.0544	0.0679	0.0396
	3	0.0551	0.2475	0.1165	0.0724	0.0839	0.0401
	4	0.0538	0.2849	0.1423	0.0863	0.1025	0.0376
	5	0.0576	0.3161	0.1653	0.0979	0.1442	0.0425
	6	0.0688	0.3513	0.1959	0.1292	0.1988	0.0630
	7	0.0821	0.3891	0.2311	0.1771	0.2532	0.0793
	8	0.0954	0.4252	0.2681	0.2376	0.3568	0.0950

5. Conclusions

The logratio transformations do not applies when compositional data have zeros. In this paper, a nonparametric method based on the multivariate Gauss kernel density estimation is suggested to deal with the rounded zeros. Because the clr coordinates add to zero, the ilr coordinates are applied in the proposed method. Under the ilr coordinates in Equation (2.2), the multivariate Gauss kernel function is related with the Aitchison distance between subcompositions. In the simulation study and real example, the proposed method is compared with the multiplicative replacement strategy, the multiplicative Kaplan-Meier method, the multiplicative lognormal replacement method, the modified EM algorithm based on alr coordinates and the robust modified EM algorithm based on ilr coordinates. The results in simulation study show that the proposed method presents a good performance in comparison with other methods in the two measures

STRESS and RDVM as the percentage of rounded zeros increases. Furthermore, in the real example, the performance of proposed method is obvious. The feature of our framework is that the proposed method works when the distribution function form is unknown. Future work will be dedicated to the study of bandwidth matrix in multivariate kernel function.

Acknowledgements

This work was supported by the Natural Science Foundation of Shanxi Province of China (No. 2015011044), Shanxi International Science and Technology Cooperation Project (No. 2015081020), and Graduate Education Innovation Project of Shanxi Province (No. 2017BY001).

References

[1] Aitchison, J. *The statistical analysis of compositional data*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1986.

[2] Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V. *Logratio analysis and compositional distance*, Mathematical Geosciences **32** (3), 271-275, 2000.

[3] Billheimer, D., Guttorp, P. and Fagan, W.F. *Statistical interpretation of species composition*, Journal of the American Statistical Association **96** (456), 1205-1214, 2001.

[4] Chacón, J.E., Mateu-Figueras, G. and Martín-Fernández, J.A. *Gaussian kernels for density estimation with compositional data*, Computers & Geosciences **37** (5), 702-711, 2011.

[5] Egozcue, J.J. and Pawlowsky-Glahn, V. *Groups of parts and their balances in compositional data analysis*, Mathematical Geosciences **37** (7), 795-828, 2005.

[6] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. *Isometric logratio transformations for compositional data analysis*, Mathematical Geosciences **35** (3), 279-300, 2003.

[7] Filzmoser, P. *StatDA: statistical analysis for environmental data, R package version 1.6.3*, 2011.

[8] Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. *Dealing with zeros and missing values in compositional data sets using nonparametric imputation*, Mathematical Geosciences **35** (3), 253-278, 2003.

[9] Martín-Fernández, J.A., Palarea-Albaladejo, J. and Olea, R.A. *Dealing with zeros*, in *Compositional Data Analysis: Theory and Applications, V. Pawlowsky-Glahn and A. Buccianti, eds.*, John Wiley & Sons Ltd., Chichester, 47-62, 2011.

[10] Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P. and Palarea-Albaladejo, J. *Model-based replacement of rounded zeros in compositional data: classical and robust approaches*, Computational Statistics & Data Analysis **56** (9), 2688-2704, 2012.

[11] Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J.J. *The normal distribution in some constrained sample spaces*, Sort-statistics and Operations Research Transactions **37** (1), 29-56, 2008.

[12] Palarea-Albaladejo J. and Martín-Fernández, J.A. *A modified EM alr-algorithm for replacing rounded zeros in compositional data sets*, Computers & Geosciences **34** (8), 902-917, 2008.

[13] Palarea-Albaladejo, J. and Martín-Fernández, J.A. *Values below detection limit in compositional chemical data*, Analytica Chimica Acta **764**, 32-43, 2013.

[14] Palarea-Albaladejo, J. and Martín-Fernández, J.A. *zCompositions-R package for multivariate imputation of left-censored data under a compositional approach*, Chemometrics and Intelligent Laboratory Systems **143**, 85-96, 2015.

[15] Palarea-Albaladejo, J., Martín-Fernández, J.A. and Gómez-García, J. *A parametric approach for dealing with compositional rounded zeros*, Mathematical Geosciences **39** (7), 625-645, 2007.

[16] Pawlowsky-Glahn, V. and Buccianti, A. *Compositional Data Analysis: Theory and Applications*, John Wiley & Sons Ltd., Chichester, 2011.

[17] Pawlowsky-Glahn, V. and Egozcue, J.J. *Geometric approach to statistical analysis on the simplex*, Stochastic Environmental Research and Risk Assessment **15** (5), 384-398, 2001.
[18] Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. *Modeling and analysis of compositional data*, Statistics in Practice, John Wiley & Sons, Ltd., Chichester, 2015.
[19] Rizzo, M.L. and Székely, G.J. *Energy: E-statistics (energy statistics), R package version 1.1-0*, 2008.
[20] Silverman, B.W. *Density estimation for statistics and data analysis*, Chapman & Hall, London, 1986.
[21] van den Boogaart, K.G. and Tolosana-Delgado, R. *Analyzing compositional data with R*, Springer, Heidelberg, 2013.