Scientific Research Publishing

Open Journal of Modelling and Simulation, 2024, 12, 33-42
https://www.scirp.org/journal/ojmsi
ISSN Online: 2327-4026
ISSN Print: 2327-4018

A Study of EM Algorithm as an Imputation Method: A Model-Based Simulation Study with Application to a Synthetic Compositional Data

Yisa Adeniyi Abolade, Yichuan Zhao

Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia, USA
Email: yabolade1@gsu.edu

How to cite this paper: Abolade, Y.A. and Zhao, Y.C. (2024) A Study of EM Algorithm as an Imputation Method: A Model-Based Simulation Study with Application to a Synthetic Compositional Data. *Open Journal of Modelling and Simulation*, **12**, 33-42.
https://doi.org/10.4236/ojmsi.2024.122002

Received: January 4, 2024
Accepted: March 5, 2024
Published: March 8, 2024

Copyright © 2024 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Compositional data, such as relative information, is a crucial aspect of machine learning and other related fields. It is typically recorded as closed data or sums to a constant, like 100%. The statistical linear model is the most used technique for identifying hidden relationships between underlying random variables of interest. However, data quality is a significant challenge in machine learning, especially when missing data is present. The linear regression model is a commonly used statistical modeling technique used in various applications to find relationships between variables of interest. When estimating linear regression parameters which are useful for things like future prediction and partial effects analysis of independent variables, maximum likelihood estimation (MLE) is the method of choice. However, many datasets contain missing observations, which can lead to costly and time-consuming data recovery. To address this issue, the expectation-maximization (EM) algorithm has been suggested as a solution for situations including missing data. The EM algorithm repeatedly finds the best estimates of parameters in statistical models that depend on variables or data that have not been observed. This is called maximum likelihood or maximum a posteriori (MAP). Using the present estimate as input, the expectation (E) step constructs a log-likelihood function. Finding the parameters that maximize the anticipated log-likelihood, as determined in the E step, is the job of the maximization (M) phase. This study looked at how well the EM algorithm worked on a made-up compositional dataset with missing observations. It used both the robust least square version and ordinary least square regression techniques. The efficacy of the EM algorithm was compared with two alternative imputation techniques, k-Nearest Neighbor (k-NN) and mean imputation ($\overline{x}$), in terms of Aitchison distances and covariance.

Keywords

Compositional Data, Linear Regression Model, Least Square Method, Robust Least Square Method, Synthetic Data, Aitchison Distance, Maximum Likelihood Estimation, Expectation-Maximization Algorithm, k-Nearest Neighbor, and Mean imputation

1. Introduction

Compositional data exclusively consists of relative information. These entities are part of a broader entity. Typically, closed data or data that aggregates to a constant value, such as 100%, is commonly documented. An illustrative instance within the field of medicine involves the examination of the constituent elements present in bodily fluids such as blood and urine. The statistical linear model is frequently employed to uncover latent associations among relevant random variables due to its user-friendly nature and interpretability. In the domain of machine learning and its associated disciplines, ensuring the quality of data is a significant difficulty. The quality of the underlying data plays a crucial role in determining the quality of information obtained through Machine Learning algorithms, as these algorithms rely solely on data for their functioning. One significant concern pertaining to data quality involves the presence of missing data, particularly in compositional datasets. The linear regression model is a widely employed statistical modeling technique that is utilized across various applications to ascertain correlations between variables of interest. The method of maximum likelihood estimation (MLE) is commonly employed to estimate the parameters of linear regression by determining the values that maximize the likelihood function given the observed data. The obtained model can be utilized for doing partial effects analysis on the independent variables as well as for making predictions about future outcomes.

However, it is important to note that many datasets often exhibit missing observations. In the context of research, it is possible for participants to opt out of providing a response to a survey query, for files to finally undergo destruction, or for data to be inadequately preserved. The process of recommencing data collecting and recovery in this case will incur financial expenses and necessitate a significant amount of time. The matter pertaining to the evaluation of inadequate data necessitates attention. The expectation-maximization (EM) technique has been proposed as a potential solution for scenarios with missing data due to its robust convergence properties. The estimation of parameters in statistical models that involve unobserved variables or unobserved data is commonly achieved by the repetitive use of the Expectation-Maximization (EM) technique, which seeks to find the maximum likelihood or maximum a posteriori (MAP) estimates. The M phase of the EM algorithm involves the estimation of parameters that maximize the expected log-likelihood obtained during the E stage. The

E-step, in contrast, constructs a function that calculates the expected value of the log-likelihood, using the current estimate as the parameters. In the subsequent E phase, these estimates are subsequently employed to determine the distribution of the latent variables or missing data.

Despite the wide investigation of the EM algorithm as an imputation tool, there is a lack of knowledge regarding its effectiveness when used for compositional data. This study investigates the performance of the EM method on a synthetic compositional dataset with missing observations. Two imputation techniques, namely the robust least square version and least square, are utilized and evaluated. The EM technique is applied to simulated studies by making iterative assumptions about a compositional dataset with random missing data and outliers, assuming a normal distribution. The effectiveness of the EM method was evaluated by comparing its results with two commonly employed imputation techniques, namely k-Nearest Neighbor (k-NN) and mean imputation ($\overline{x}$), in terms of Aitchison distance and covariance [1]. Based on the conducted trials, it was shown that the robust variant of the EM algorithm exhibited superior performance compared to alternative imputation strategies.

2. Methodolgy

2.1. Linear Regression Model

We take into consideration a one-dimensional estimator and response linear regression model. Let's say we have n observations in our dataset. We define the predictor $X=\left(x_1,x_2,\cdots,x_n\right)$, and the response $Y=\left(y_1,y_2,\cdots,y_n\right)$. For the i^{th} observation, we assume that y_i and x_i are related by the linear regression model and in Equation (1):

$$y_i=\beta_0+\beta_1 x_i+\epsilon,\ \epsilon\sim NID\left(0,\sigma^2\right) \tag{1}$$

We assume that $x_i\sim N\left(\alpha,\delta^2\right)$, *i.i.d.* Under such assumptions, the conditional distribution of Y given X is $\left[Y\mid X\right]\sim N\left(\beta_0+\beta_1 X,\sigma^2\right)$. Then we can write down the joint probability density of X and Y given by

$$\begin{aligned} f\left(y_i,x_i\right)&=f\left(y_i\mid x_i\right)f\left(x_i\right)\\ &=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2\sigma^2}\left(y_i-\beta_0-\beta_1 x_i\right)^2}\times\frac{1}{\delta\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2\delta^2}\left(x_i-\alpha\right)^2} \end{aligned} \tag{2}$$

2.2. Missing Values

The data is not completely observed in a lot of real-world scenarios. We expand our model to have response values Y fully observed and only m of the predictor values observed (*i.e.*, $n-m$ predictor values missing), and the response values Y are fully observed. We can arrange the dataset so that the first m observation is fully observed.

$$X_{comp}=\left(x_1,\cdots,x_m,x_{m+1},\cdots,x_n\right)=\left(X_{obs},X_{miss}\right) \tag{3}$$

Equation (4) thus allows for the decomposition of the entire data log-likelihood for the model into the observable and missing parts.

$$\begin{aligned} L(\theta;X,Y) &= s_{i=1}^{n} L(\theta;x_i,y_i) \\ &= -2n\log\sqrt{\sigma^2 2\pi} - 2n\log\sqrt{\delta^2 2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1 x_i)^2 \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=m+1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 - \frac{1}{2\delta^2}\sum_{i=1}^{m}(x_i-\alpha)^2 \\ &\quad -\frac{1}{2\delta^2}\sum_{i=m+1}^{n}(x_i-\alpha)^2 \end{aligned} \tag{4}$$

where $\theta=(\beta_0,\beta_1,\sigma^2,\alpha,\delta^2)\in\mathbb{R}^5$

2.3. The EM Algorithm Formulation

The issue with the above calculation is that X_{mis} with not observed and needs to be estimated. One reasonable approach is that we simply require each $x_{m+1},\cdots,x_n$ to be replaced by its conditional expectation given the observed data, X_{obs} and Y.

1) *E-step*:

$$\begin{aligned} &E\left[\sum_{i=m+1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right] \\ &=\sum_{i=m+1}^{n}\left((y_i-\beta_0)^2+\beta_1^2 E(X_i^2\mid y_i,\theta^*)-2\beta_1(y_i-\beta_0)E(X_i\mid y_i,\theta^*)\right) \\ &\quad E\left[\sum_{i=m+1}^{n}(x_i-\alpha)^2\right] \\ &\quad =\sum_{i=m+1}^{n}\left(E(X_i^2\mid y_i,\theta^*)-2\alpha E(X_i\mid y_i,\theta^*)+\alpha^2\right) \end{aligned} \tag{5}$$

where $E[x_i\mid y_i,\theta^*]$ and $E[x_i^2\mid y_i,\theta^*]$ are the first and second conditional moments, respectively. Since X and Y have a bivariate normal distribution, we can derive the conditional of X given Y and θ^*

$$E[X_i\mid y_i,\theta^*]\sim N\left(\alpha+\frac{\beta_1\delta^2}{\sigma^2+\beta_1^2\delta^2}(y_i-\beta_0-\beta_1\alpha),\frac{\sigma^2\delta^2}{\sigma^2+\beta_1^2\delta^2}\right) \tag{6}$$

Then we can easily find the conditional first and second moment of X_{miss} given Y and θ^*, denoted as M^1 and M^2 respectively.

$$M_i^1=\alpha+\frac{\beta_1\delta^2}{\sigma^2+\beta_1^2\delta^2}(y_i-\beta_0-\beta_1\alpha) \tag{7}$$

$$M_i^2=\left(\alpha+\frac{\beta_1\delta^2}{\sigma^2+\beta_1^2\delta^2}(y_i-\beta_0-\beta_1\alpha)\right)^2+\frac{\sigma^2\delta^2}{\sigma^2+\beta_1^2\delta^2} \tag{8}$$

With these terms computed above, the E-step formulation is shown in Equation (9) below.

$$\begin{aligned} Q\left(\theta,\theta^*\right) &= L\left(\theta;X,Y\right) \\ &= -2n\log\sqrt{2\pi} - n\log\delta - n\log\sigma \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(\left(y_i-\beta_0\right)^2+\beta_1^2M^2-2\beta_1\left(y_i-\beta_0\right)M^1\right) \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=m+1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2-\frac{1}{2\delta^2}\sum_{i=m+1}^{n}\left(x_i-\alpha\right)^2 \\ &\quad -\frac{1}{2\delta^2}\sum_{i=1}^{m}\left(M^2-2\alpha M^1-\alpha^2\right) \end{aligned} \tag{9}$$

where M^1, M^2 are given in Equations (7) and (8).

2) *M-step*:

The M-step maximizes $Q\left(\theta,\theta^*\right)$ calculated in the E-step. Solving $\frac{\partial Q\left(\theta,\theta^*\right)}{\partial\theta}=0$, we get the following results. The updated estimates of β' is just the OLS solution to the model, *i.e.*, $\beta'=\left(X^{\mathrm{T}}X\right)^{-1}\left(X^{\mathrm{T}}Y\right)$

$$\begin{bmatrix}\beta_0'\\ \beta_1'\end{bmatrix}=\begin{bmatrix}n & \sum_{i=1}^{n}\widetilde{x_i^*}\\ \sum_{i=1}^{n}\widetilde{x_i^*} & \sum_{i=1}^{n}\widetilde{x_i^{2^*}}\end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^{n}y_i\\ \sum_{i=1}^{n}\widetilde{x_i^*}*y_i\end{bmatrix}$$

where $\widetilde{X^*}=\left(X_{obs},M^{1*}\right)\in\mathbb{R}^n$ and $\widetilde{X^{2^*}}=\left(X_{obs}^2,M^{2*}\right)\in\mathbb{R}^n$ are estimated completed predictor under current estimated parameter θ^*. Similarly, the other updated parameters will be:

$$\sigma^{2'}=\frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i-\beta_0'\right)-2\beta_1'y_i-\beta_0'\right)\widetilde{x_i^*}+\left(\beta_1'\right)^2\widetilde{x_i^{2^*}}$$

$$\alpha'=\frac{1}{n}\sum_{i=1}^{n}\widetilde{x_i^*}$$

$$\delta^{2'}=\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{x_i^{2^*}}-2\alpha'\widetilde{x_i^*}+\alpha'^2\right)$$

2.4. Convergence of EM Algorithm

We will discuss the convergence of EM algorithm in a more general setting. Suppose we have dataset of m independent examples and want to fit a parametric model $p\left(x,z\right)$ to the dataset, then the log likelihood function is:

$$\begin{aligned} l\left(\theta\right) &= \sum_{i=1}^{m}\log p\left(x^{(i)};\theta\right) \\ &= \sum_{i=1}^{m}\log p\left(x^{(i)},z;\theta\right) \end{aligned}$$

where z are the latent random variables. Explicitly finding the maximum likelihood estimate of the parameter θ is quite hard, but if $z^{(i)}$ is observed, the estimation would be easy. Let $Q_i\left(z\right)\geq 0$, $\sum_z Q_i\left(z\right)=1$. Since $f\left(x\right)=\log\left(x\right)$ is a concave function and by Jensen's Inequality, we get the lower-bound of $l\left(\theta\right)$:

$$\sum_i\log p\left(x^{(i)};\theta\right)=\sum_i\sum_{z^{(i)}}p\left(x^{(i)},z^{(i)};\theta\right) \tag{10}$$

$$= \sum_i \log \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \tag{11}$$

$$\geq \sum_i \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \tag{12}$$

Note that this inequality holds for any distribution Q_i, it gives the lower bound on $l(\theta)$. Later we will show that the $l(\theta)$ increases monotonically with successive iterations of EM if the lower-bound is tight at θ. We know that the Jensen's Inequality holds with equality if the random variables are constant. So, it suffices to satisfy:

$$\frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} = c$$

$$i.e.\ Q_i\left(z^{(i)}\right) \propto p\left(x^{(i)}, z^{(i)}; \theta\right)$$

where c is constant and does not depend on $z^{(i)}$. Under this assumption, since $\sum_z Q_i(z) = 1$, we get:

$$Q_i\left(z^{(i)}\right) = \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{\sum_z p\left(x^{(i)}, z^{(i)}; \theta\right)} = \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{p\left(x^{(i)}, z; \theta\right)} = p\left(z^{(i)} \mid x^{(i)}; \theta\right)$$

So Q_i*'s* is just the posterior distribution of the $z^{(i)}$'s given $x^{(i)}$ and the setting of θ. Now we introduce the iteration of EM algorithm:

While $\left\|\theta^{(t)} - \theta^{(t-1)}\right\| > \epsilon$ do

E-step:

Compute $Q_i^{(t)}\left(z^{(i)}\right) := p\left(z^{(i)} \mid x^{(i)}; \theta^{(i)}\right)$

M-step

Compute

$$\theta^{(t+1)} := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i^{(t)}\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i^{(t)}\left(z^{(i)}\right)}$$

End

Consider that:

$$l\left(\theta^{(t+1)}\right) \geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta^{(t+1)}\right)}{Q_i^{(t)}\left(z^{(i)}\right)} \tag{13}$$

$$\geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta^{(t)}\right)}{Q_i^{(t)}\left(z^{(i)}\right)} \tag{14}$$

$$= l\left(\theta^{(t)}\right) \tag{15}$$

The first inequality holds because of (13). By the definition of $\theta^{(t+1)}$, (15) is obvious. The last equality holds because lower-bound in (13) is tight at $\theta = \theta^{(t)}$ under our previous assumption. So, the sequence $\left\{l\left(\theta^{(t)}\right)\right\}_t$ is both upper

bounded (by 0) and increasing. Hence, in EM algorithm the log likelihood converges monotonically.

3. Application

This section covers the model-based simulation research application of the EM algorithm with least square and resilient least square regression on compositional data. We also analyze the resilience and efficiency of the EM approach and compare its output to two other commonly used imputation techniques, k-Nearest Neighbor (k-NN) and mean imputation ($\overline{x}$), to address missing data.

3.1. Data Description

Compositional data is a unique kind of non-negative data that contains the pertinent information not in the actual data values but rather in the ratios between the variables. An observation $x=\left(x_1,\cdots,x_D\right)$ is a D-part composition if, and only if, $x_i>0$, $i=1,\cdots,D$, and according to Aitchison [2], the ratios between the components include all the important information.

$S^D=\left\{\left[x_1,\cdots,x_D\right]:x_i>0\left(i=1,\cdots,D\right),x_1+\cdots+x_D=1\right\}$ and $\left(x_1,\cdots,x_D\right)=\dfrac{\left(w_1,\cdots,w_D\right)}{w_1+\cdots+w_D}$ where (w) denote = total weight and $\left(w_1,\cdots,w_D\right)$ are the component weights. According to Aitchison [2], compositional data is not directly represented in Euclidean space. The Aitchison distance d_A is a suitable way to measure the D-part composition known as (simplex) [3]. According to Egozcue *et al.* [4], the isometric log-ratio (*ilr*) is used to convert the D-dimensional simplex into the real space $\mathbb{R}^{D-1}$. With this transformation, the Aitchison distance can be expressed as $d_A\left(x,y\right)=d_E\left(ilr\left(x\right),ilr\left(y\right)\right)$, where d_E denotes the Euclidean distance. The data in this simulation is generated by a normal distribution on the simplex, denoted by $\mathbb{N}_s^D\left(\mu,\Sigma\right)$ (Mateu-Figueras, Pawlowsky-Glahn, and Egozcue) [5]. We generated 10,000 realizations of a random variable [6] $X\sim\mathbb{N}_s^4\left(\mu,\Sigma\right)$ with $\mu=\left(0,2,3\right)^{\mathrm{T}}$ and $\Sigma=\left(\left(1,-0.5,1.4\right)^{\mathrm{T}},\left(-0.5,1,-0.6\right)^{\mathrm{T}},\left(1.4,-0.6,2\right)^{\mathrm{T}}\right)$ [7].

3.2. Experimental Design

- To assess the effectives of the EM algorithm, we look at the results using least squares (LS) and its robust version (RLS) across a range of missing data rates (contamination levels between 5% and 10%) and outlier rates (1%, 3%, 5%, and 10%) expressed in terms of their Aitchison distance (d_A) [8].
- We also look at the EM algorithm's output in terms of the covariances of various rates of outliers (1%, 3%, 5%, and 10%) and missing data (contamination levels ranging from 5% to 10%).

3.3. Results and Analysis

The detailed results for all experiments are discussed in this section (**Table 1**, **Table 2**, **Figure 1**, **Figure 2**).

Table 1. Performance metrics for different imputation methods in terms of distance.

S/N	Epsilon[1]	NArate[2]	xMean[3]	kNN[4]	LS[5]	RLS[6]
1	0.01	0.05	1.208117	0.8815717	0.9277800	**0.7544389**
2	0.03	0.05	1.869092	1.6451996	1.9300590	**1.5242680**
3	0.05	0.05	2.613898	2.2722186	2.5338586	**2.1552349**
4	0.10	0.05	4.044281	3.8628814	4.1707580	**3.6228353**
5	0.01	0.10	1.087026	0.7727159	0.8874163	**0.6802822**
6	0.03	0.10	1.419570	1.2479738	1.3840970	**1.0872886**
7	0.05	0.10	1.798377	1.5482930	1.7860411	**1.4121292**
8	0.10	0.10	2.693001	2.4922940	2.8084783	**2.2551779**

[1]Epsilon denote rate of outlier at (1%, 3%, 5% and 10%); [2]NArate denote missing rate (contamination level at 5% and 10%); [3]xMean denote arithmetic mean imputation method; [4]kNN denote k-Nearest Neighbor imputation method; [5]LS denote least square regression version of EM algorithm; [6]RLS denote robust least square regression version of EM algorithm; *Bold numbers indicate the best performing imputation method for a given epsilon and missing rate.

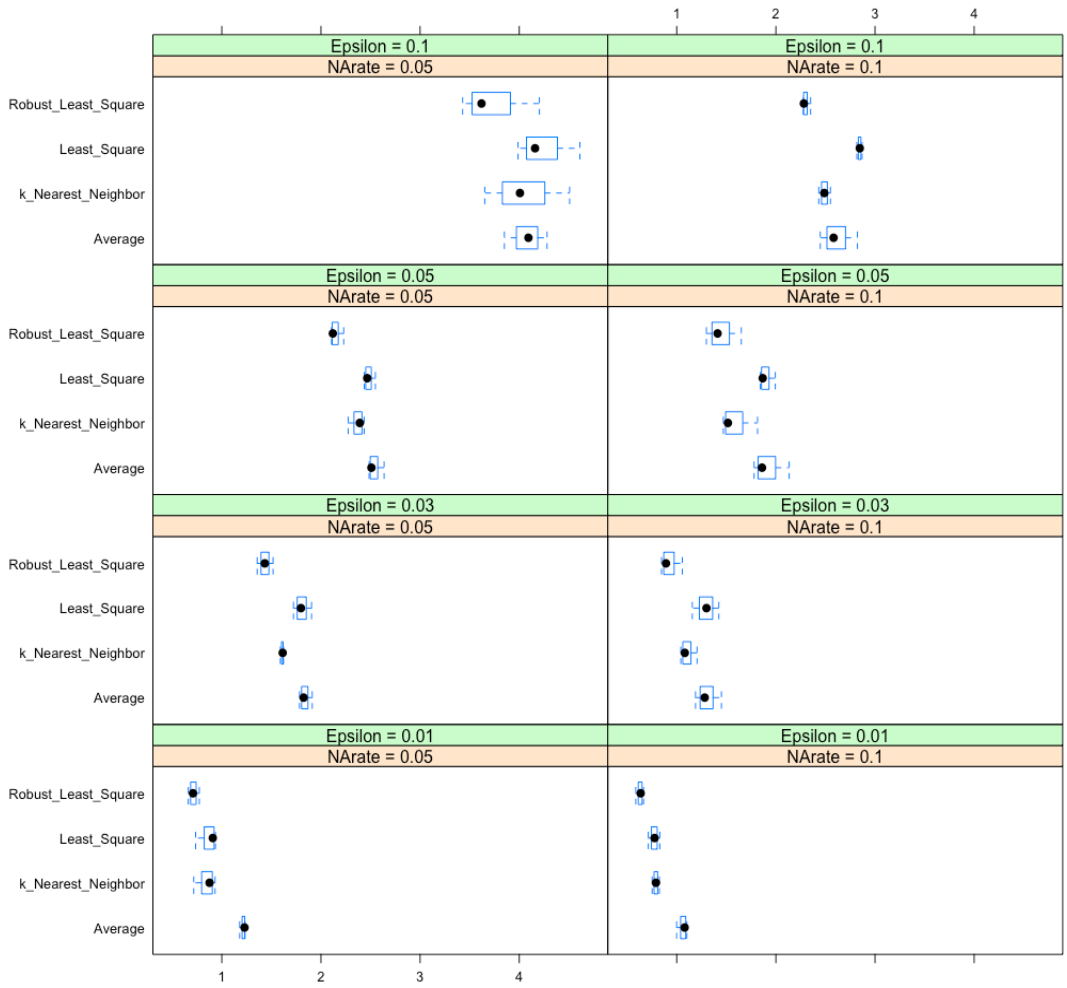

Figure 1. Performance for different imputation methods in terms of distance.

Table 2. Performance metrics for different imputation methods in terms of Covariance.

S/N	Epsilon[1]	NArate[2]	xMean[3]	kNN[4]	LS[5]	RLS[6]
1	0.01	0.05	0.1982435	0.1686442	0.1756952	**0.1668150**
2	0.03	0.05	0.4107212	0.4087216	0.3966203	**0.3836261**
3	0.05	0.05	0.5342290	0.5234262	0.5219798	**0.5126078**
4	0.10	0.05	0.7064511	0.7005467	0.7037723	**0.6945066**
5	0.01	0.10	0.2274753	0.1765070	0.1968090	**0.1695136**
6	0.03	0.10	0.4497979	0.3934609	0.4346386	**0.3849345**
7	0.05	0.10	0.5563971	0.5472681	0.5541094	**0.5150961**
8	0.10	0.10	0.7163915	0.6997279	0.7142401	**0.6907288**

[1]Epsilon denote rate of outlier at (1%, 3%, 5% and 10%); [2]NArate denote missing rate (contamination level at 5% and 10%); [3]xMean denote arithmetic mean imputation method; [4]kNN denote k-Nearest Neighbor imputation method; [5]LS denote least square regression version of EM algorithm; [6]RLS denote robust least square regression version of EM algorithm; *Bold numbers indicate the best performing imputation method for a given epsilon and missing rate.

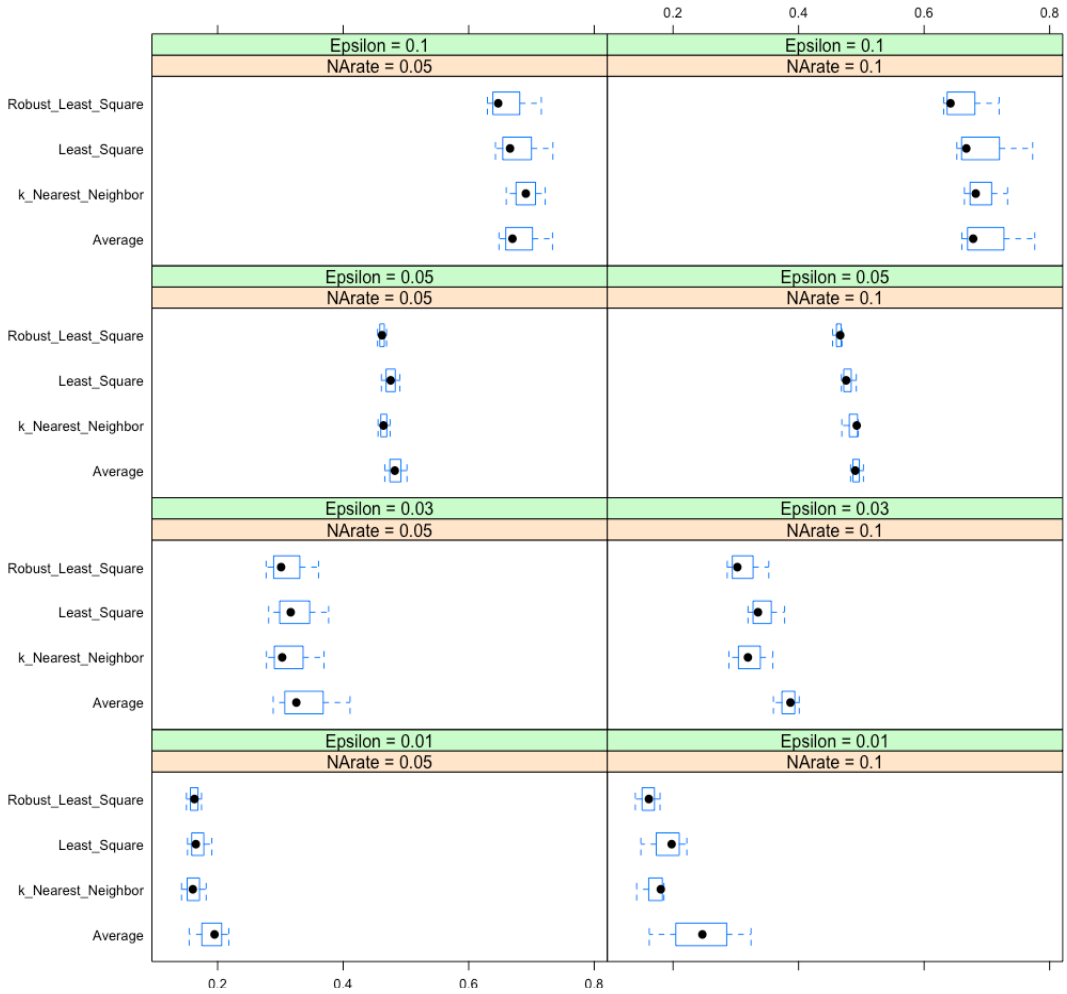

Figure 2. Performance for different imputation methods in terms of Covariance.

4. Conclusions

In this study, it is postulated that the missing rate follows a random pattern. Consequently, both the least square and robust least square EM algorithm approaches are employed to address the issue of missing compositional data. The simulated compositional synthetic data was utilized as the practical dataset in our study. It has been shown that when the occurrence of missing and outlier data decreases, the Aitchison distance approaches zero. Furthermore, the robust least square variant of the EM method yields the smallest values in this regard making it the best performing algorithm. Distance values close to zero represent the most efficacious imputation strategies.

We also compare the Expectation-Maximization (EM) method to arithmetic mean imputation and k-Nearest Neighbor and find that the robust EM version is the best of the bunch when it comes to co-variances. In this study, we provided evidence to support the effectiveness of all strategies when applied under conditions of low missing rates, specifically those below 5%. The EM algorithm demonstrates improved performance as the rate of missing data increases. However, when the incidence of missing data exceeds 10%, most imputation methods become inadequate in generating a dependable approximation, hence exacerbating the difficulty in recovering the data.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Aitchison, J. (2002) Simplicial Inference. In: Viana, M.A.G. and Richards, D.S.P., Eds., *Contemporary Mathematics Series*, *Vol.* 287: *Algebraic Methods in Statistics and Probability*, American Mathematical Society, Providence, 1-22.

[2] Aitchison, J. (1986) The Statistical Analysis of Compositional Data. Chapman & Hall, London.

[3] Aitchison, J. (1989) Measures of Location of Compositional Data Sets. *Mathematical Geology*, **21**, 787-790. https://doi.org/10.1007/BF00893322

[4] Martín-Fernández, J.A., Egozcue, J.J., Olea, R.A. *et al.* (2021) Units Recovery Methodsin Compositional Data Analysis. *Natural Resources Research*, **30**, 3045-3058. https://doi.org/10.1007/s11053-020-09659-7

[5] Pawlowsky, V., Olea, R.A. and Davis, J.C. (1995) Estimation of Regionalized Compositions: A Comparison of Three Methods. *Mathematical Geosciences*, **27**, 105-127. https://doi.org/10.1007/BF02083570

[6] Weltje, G.J. (1997) End-Member Modeling of Compositional Data: Numerical-Statistical Algorithms for Solving the Explicit Mixing Problem. *Mathematical Geosciences*, **29**, 503-549. https://doi.org/10.1007/BF02775085

[7] Rehder, U. and Zier, S. (2001) Comment on "Logratio Analysis and Compositional distance by Aitchison *et al.* (2000)". *Journal of Mathematical Geology*, **32**, 741-763.

[8] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**, 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x