MJFAS
Malaysian Journal of Fundamental and Applied Sciences

RESEARCH ARTICLE

Imputation Method based on Adaptive Group Lasso for High-dimensional Compositional Data with Missing Values

Ying Tian[a], Majid Khan Majahar Ali[a*], Lili Wu[a,b], Tao Li[a,c]

[a]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Gelugor, Pulau Pinang, Malaysia; [b]Department of Computer Science, Xinzhou Teachers University, 034000 Xinzhou, Shanxi, China; [c]Weifang University of Science and Technology, 262713 Shouguang, Shandong, China

Abstract Compositional data usually refers to data on the individual components that make up a whole. Such data are common in many fields, especially in chemistry, biology, geology, and other scientific and engineering fields. However, in many real-life situations, a large number of missing values are often collected. The complexity of compositional data with missing values makes traditional estimation methods seem overwhelming. Therefore, how to effectively perform statistical inference on compositional data with missing values has attracted the attention of many scholars in recent years. The logarithmic scale transformation provides a possibility for compositional data, but this transformation has limited requirements on the components of the compositional data, such as not including or missing fixes and constraints. Therefore, it is of great significance to explore a new estimation method for composition data theoretically. In this paper, a compositional data imputation method based on the adaptive group least absolute shrinkage and selection operator (AGLasso) is proposed. AGLasso is able to adapt imputation methods to different data distributions and patterns based on the characteristics of the data. While traditional methods may result in lost or biased information, AGLasso attempts to impute while preserving data integrity. Through data analysis, the imputation effect of compositional data containing missing values is compared under different missing rates and correlation coefficients, and a comparative study is conducted with the Lasso imputation method, the adaptive Lasso imputation method and the group Lasso imputation method. The results show that adaptive group Lasso is superior to the other three interpolation methods. In domains such as healthcare data, where data quality has a huge impact on decision making, AGLasso can help improve data integrity and usability. And in the future, Generative Adversarial Network (GAN)-based imputation methods and novel deep learning methods using techniques such as self-encoders are expected to show more power in dealing with missing values.

Keywords: Compositional data, imputation method, machine learning, adaptive group lasso.

***For correspondence:**
majidkhanmajaharali@usm.my

Received: 03 June 2024
Accepted: 16 Oct. 2024

Introduction

Compositional data is data consisting of components (or constituents), where each component represents a part of the whole and the sum of all components constitutes the whole. This type of data is usually expressed as relative proportions or percentages rather than absolute quantities [36]. However, most statistical analysis methods are based on complete data, and the log-ratio transformation will not be implemented when there are missing values in the data set, so the treatment of missing values in high-dimensional compositional data is of great significance [8].

For the problem of imputation of missing values in high dimensional compositional data, common approaches include, mean or median imputation: for each missing component, imputation is performed using the mean or median value that has been observed for that component. The k-nearest neighbor imputation method utilizes the idea of k-nearest neighbors to find samples that are similar to the pattern

e-ISSN 2289-599X | DOI: https://doi.org/10.11113/mjfas.v20n3.3331

of the missing component and then interpolates using the component values of those samples [44]. Model imputation methods utilize information from other component and non-component data to build an appropriate model and then use that model to interpolate the missing values [3][4]. When choosing an imputation method, the nature of the data, the relationships between the components, the information available, and the assumptions of the interpolation method need to be considered. The choice of imputation method may also involve model complexity and computational complexity. In practice, it is often necessary to experiment and compare methods on a case-by-case basis to find the most appropriate method for the data set [19].

Missing data refers to instances where certain observations in a dataset are absent or not recorded. These gaps may arise due to technical issues, oversights during data collection, or intentional data deletion. Based on the mechanism of the missingness, missing data is generally categorized into three types: Missing Completely at Random (MCAR), where the missingness is unrelated to any variables; Missing at Random (MAR), where the missingness is related to some observed variables; and Missing Not at Random (MNAR), where the missingness is related to the unobserved values themselves. Understanding the type of missing data is crucial for selecting the appropriate handling method. In practice, common imputation methods are single imputation [31], multiple imputation [33], etc., these imputation methods have more or less shortcomings [1]. For example, mean imputation [29], in practice, easy to underestimate the variance and the accuracy of regression interpolation strongly depends on the quality of auxiliary information. Therefore. The selection of the missing value treatment method is particularly important. If they are not handled properly, the results of the statistical analysis will have a serious negative impact [9].

In imputation methods, high-dimensional data often contain a lot of redundancy and noise, which increases the complexity and computational cost of the imputation model. For statistical analysis of high-dimensional data, it is usually necessary to consider the variable selection strategy for dimensionality reduction, and some previously proposed dimensionality reduction methods, such as clustering, partial least squares. The processing results of principal component regression, ridge regression, and tree-based integration methods are not ideal [15]. Through dimensionality reduction techniques, the data dimensionality can be reduced and the most informative features can be extracted, thus simplifying the interpolation process and improving the accuracy and efficiency of the imputation. This process not only reduces the computational burden, but also reduces the risk of model overfitting.

The problem is the dimensionality reduction technique for high-dimensional compositional data. Due to the existence of these characteristics of high-dimensional data: dimensional catastrophe, overfitting computational complexity, feature selection, data sparsity, etc., which makes it difficult to analyze and process high-dimensional data. In this case, the processing methods that have been successfully applied to data in low-dimensional space are no longer adapted to high-dimensional data, which has led to the creation of some new methods for high-dimensional data [38]. Data dimensionality reduction is a particularly effective method for solving the difficulties of analyzing and processing high-dimensional data, and has been widely applied in the fields of data compression, data mining, machine learning, pattern recognition, and visualization of data.

Lasso stands for "Least absolute shrinkage and selection operator." It is a statistical method used for variable selection and regularization in linear regression models [41]. Lasso adds a penalty term to the ordinary least squares (OLS) objective function, which includes the sum of squared residuals. The penalty term is proportional to the absolute values of the regression coefficients. This paper employs the Adaptive Group Lasso algorithm to address the imputation problem in high-dimensional compositional data with missing values. The Adaptive Group Lasso (AGLasso) algorithm is a variable selection method that effectively addresses the limitations of traditional imputation methods in terms of estimation accuracy and computational speed. AGLasso is an extension of the standard Lasso and Group Lasso methods. It was designed to address some of the limitations inherent in these earlier techniques, particularly in the context of handling grouped variables and achieving more accurate variable selection. It introduces an adaptive weighting mechanism that assigns different penalties to different groups based on their importance. This is typically achieved by using weights inversely proportional to some initial estimates of the coefficients, which can more effectively differentiate between relevant and irrelevant groups of variables. This leads to better variable selection, particularly in high-dimensional settings where many variables are present but only a few are truly influential. On one hand, compared to the model parameters of traditional imputation methods, Adaptive Group Lasso can simultaneously perform parameter estimation and variable selection, thus enhancing the accuracy of imputation estimates. On the other hand, Adaptive Group Lasso is an extension of Group Lasso and Adaptive Lasso, combining the advantages of both [16].

To enhance the computational speed of Adaptive Group Lasso, we transform it into a two-step Group Lasso solution. This algorithm not only relaxes the assumptions of the model but also significantly improves the solving speed of Group Lasso, providing great convenience for handling high-dimensional data. In this study, we compare this method with Adaptive Lasso, Group Lasso, and Lasso, validating the superiority of the new method in terms of missing rates and correlation coefficients [27].

AGLasso further considers the correlation of variable groups on the basis of Adaptive Lasso. By weighting the variable groups, it is able to deal with the intrinsic structure and correlation in the data, and is suitable for interpolation tasks where there is significant correlation between variables in high-dimensional data. The new AGLasso estimation method was compared with ALasso, GLasso and Lasso imputation methods for compositional data with missing values. The performance of the proposed model is quantitatively evaluated using MSE, MADE, RMSE and NRMSE. In the AGLasso method, the minimum 3.12% and maximum 12.68% percentage improvement in MSE values; the minimum 5.70% and maximum 18.27% percentage improvement in MADE values; the minimum 6.22% and maximum 26.60% percentage improvement in RMSE values; the minimum 4.27% and maximum 14.83% percentage improvement in NRMSE values. Based on the results, it can be concluded that the AGLasso imputation method outperforms some existing methods in prediction accuracy and variable selection. Therefore, this paper not only identifies a solution for imputation problems in compositional data that ensures estimation accuracy while achieving faster computational speed but also expands the application scope of AGLasso. It enriches the application of Lasso theory in the field of imputation problems. Combining AGLasso with deep learning can significantly enhance missing value imputation. Deep learning's capabilities in feature extraction and nonlinear modeling can optimize the adaptive weights of AGLASSO, thereby improving imputation accuracy. This integration not only addresses complex data patterns but also tackles high-dimensional data challenges, extending applications to fields such as medical imaging and financial data.

The model evaluations used in this study are Mean Squaree Error (MSE), Mean Aitchison Distance Error (MADE), Root Mean Square Error (RMSE) and Normalized Root Mean Square Error (NRMSE). Evaluation model metrics are needed to determine the quality and accuracy of the model. The purpose of filling in missing values in compositional data is to make the filled-in data as close as possible to the actual complete data. This closeness is reflected in the true value versus the filled value. Therefore, we can judge the estimation effect of this series of filling methods by the degree of this proximity.

Materials and Methods

The origin of the Adaptive Group Lasso (AGLasso) imputation method can be traced back to the Lasso imputation method, and its development has undergone a series of improvements and extensions from the Lasso imputation method to the Adaptive Lasso (ALasso) imputation method, then to the Group Lasso (GLasso) imputation method, and finally to the Adaptive Group Lasso imputation method.

Lasso Imputation Method

The origin of The Lasso method was initially proposed by Tibshirani [34]. Its main idea is to introduce a penalty term on top of the least squares estimation by enforcing the sum of the absolute values of the parameters to be less than a constant, thereby minimizing the sum of squared residuals. The Lasso algorithm for solving the imputation problem for datasets containing missing values can be expressed as follows [6]:

$$\hat{\beta}^{Lasso}(\lambda) = arg\ min\left\{\|Y - X\beta\|_2^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{1}$$

where the regularization parameter $\lambda > 0$ is a norm for the vector $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T, (i = 1,2, \cdots, n)$is the i-th $p \times 1$explanatory variable, and the vector $Y = (y_1, y_2, \cdots, y_n)^T$is a a vector of $n \times 1$ explained variable, $X = (x_1, x_2, \cdots, x_n)^T$ is the matrix of $n \times p$ explanatory variables, and $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is the coefficients vector of the $p \times 1$ explanatory variables [14].

According to Equation (1), it can be observed that the main idea of the Lasso method is to introduce a penalty term on top of the least squares estimation. This penalty term compresses some model coefficients, driving certain coefficients to be zero, ultimately achieving variable selection. By incorporating this penalty term, Lasso enhances the sparsity of the model, making it advantageous for handling high-dimensional linear regression problems [35].

The selection of the penalty parameter λ is a crucial step in Lasso estimation. A larger penalty parameter leads to greater punishment on coefficients, making them more likely to be compressed to zero. Consequently, the entire model becomes more sparse, which may result in overlooking some important variables, causing the model to exhibit significant bias. When the regularization parameter λ is too small, retaining too many explanatory variables may lead to a highly redundant model. This can diminish the model's explanatory power and increase the risk of overfitting.

Therefore, selecting an appropriate penalty parameter is essential. A commonly used approach is cross-validation, which involves randomly dividing the original sample into a training set and a validation set. Initially, the model is trained using the training set, and its performance is then evaluated on the validation set. This process is repeated, and the model with the optimal performance is selected. In this study, K-fold cross-validation is employed.

Since Tibshirani first introduced Lasso in 1996, the method has established a theoretical foundation for handling missing data by incorporating L_1 norm constraints for variable selection and regularization. In 2006, Chen and Ibrahim [7] were the first to apply Lasso directly to missing data, exploring its effectiveness in variable selection. Zou [45] later advanced the method by introducing adaptive Lasso, which enhances its capability to handle complex data by applying varying penalties to different coefficients. In 2010, Suvrit Sra and colleagues [10][23] proposed a Lasso-based matrix completion method, particularly suited for addressing missing values in large-scale datasets. In 2012, Wainwright and Loh [21] extended Lasso's applicability to high-dimensional data by studying methods for missing data imputation in such contexts. In 2013, Simon and Friedman [32] introduced Adaptive Group LASSO, merging Group LASSO and Adaptive LASSO principles to enable more flexible variable selection. In 2018, Zhao and Rocha [43] applied Adaptive Group LASSO to missing data for the first time, developing an algorithm that effectively addresses complex missing patterns in high-dimensional datasets. Moving into the 2020s, Lasso was further refined by integrating deep learning techniques, such as autoencoders [28], significantly improving interpolation performance under complex missing data patterns.

Adaptive Lasso Imputation Method

The Lasso method has a drawback in that it applies the same level of compression to all coefficients, making it prone to excessive shrinkage for coefficients with larger absolute values. This can lead to overly sparse results. Recognizing this issue, an improvement to the Lasso method, introducing the Adaptive Lasso imputation method [20].

The Adaptive Lasso method, proposed by [18], is an improvement upon the Lasso method. In the Lasso method, each coefficient experiences the same level of shrinkage. On the other hand, the Adaptive Lasso method introduces weighting to the coefficients in the penalty term. This results in varying degrees of compression for different coefficients, where larger coefficients receive a smaller penalty, and smaller coefficients receive a larger penalty. This adaptive adjustment of penalties for different coefficients helps achieve consistent estimation results. The expression for the Adaptive Lasso method is as follows [30]:

$$\hat{\beta}^{ALasso}(\lambda) = arg\ min\left\{\|Y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \widehat{\omega}_j \left|\beta_j\right|\right\} \tag{2}$$

where λ_n is the adjustment parameter used to balance the penalty term and empirical risk. $\widehat{\omega} = 1/\left|\hat{\beta}\right|^v, v > 0$ is the adaptive penalty weight, which works by making the penalty less for the more important variables. The penalty term expression is $\lambda_n \sum_{j=1}^{p} \widehat{\omega}_j \left|\beta_j\right|$, the regression coefficient β obtained with L_1 parametric will have less non-zero components and get more sparse solutions, so L_1 parametric number can be used for feature selection [37].

Group Lasso Imputation Method

The Group Lasso method is another significant improvement upon the Lasso method [24]. Unlike the Lasso method, the Group Lasso method groups coefficients together and performs variable selection based on the level of each group. Consider the following generalized linear regression model with L factors:

$$Y = \sum_{l=1}^{L} X_l \beta_l + \varepsilon \tag{3}$$

where β_l is a p_l-dimensional coefficients corresponding to the l-th factor. According to Group Lasso's "integer-in-integer-out" property, if a group is selected, then all coefficients in the group have non-zero estimates, and therefore the corresponding variable type is considered significant, otherwise all coefficients of the variables included in that variable type are zero. Extending the Lasso method to groups gives rise to Group Lasso:

$$\hat{\beta}^{GLasso}(\lambda) = arg\ min\left\{\left\|Y - \sum_{l=1}^{L} X_l\beta_l\right\|_2^2 + \lambda\sum_{l=1}^{L}\|\beta_l\|_2\right\} \tag{4}$$

Adaptive Group Lasso Imputation Method

The Adaptive Group Lasso method (AGLasso), employed in this study, is a further extension of the Group Lasso method, integrating the advantages of the Adaptive Lasso method [25]. While the Group Lasso method applies the same weight to each group of coefficients, leading to uniform shrinkage for each group, it may result in excessive compression for group vectors with larger norms. The Adaptive Group Lasso method addresses this issue by incorporating the benefits of the Adaptive Lasso method. It introduces weighting to each group of coefficients in the penalty term, ensuring varying degrees of shrinkage for each group. This allows for a milder penalty on group vectors with larger norms and a stronger penalty on group vectors with smaller norms. By performing variable selection at the group level and adaptively adjusting the compression for each group, it enhances the accuracy of variable selection, the following Adaptive Group Lasso (AGLasso) method [11]:

$$\hat{\beta}^{AGLasso}(\lambda) = arg\ min\left\{\left\|Y - \sum_{l=1}^{L} X_l\beta_l\right\|_2^2 + \sum_{l=1}^{L}\lambda_l\|\beta_l\|_2\right\} \tag{5}$$

where $\lambda_l = \lambda\left\|\hat{\beta}_l\right\|^{-v}$,$\hat{\beta}_l$ is the consistent estimator without penalty term. Considering that this paper is a high-dimensional linear regression model, the traditional least-squares method can not get the corresponding results [36].

$$\text{Let } X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(D)} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nD} \end{pmatrix} \tag{6}$$

be the compositional data matrix. The compositional data with missing values, an iterative algorithm via Adaptive Group Lasso imputation method can be summarized as follows [40].

Step 1: Isometric logarithmic ratio transformation
Let $ilr(x_i) = z_i = (z_{i1}, z_{i2}, \cdots, z_{ip})^T, (p = D - 1; i = 1,2,\cdots,n)$, the compositional data in simplex space are transformed into real numbers in Euclidean space, the transformation here is called isometric log-ratio transformations (ilr):

$$z_i = \sqrt{\frac{D-i}{D-i+1}}\, ln\frac{\sqrt[D-i]{\prod_{l=i+1}^{D} x_l}}{x_i}, i = 1,2,\cdots,d. \tag{7}$$

Step 2: Feature grouping
According to China's resident income data classification policy, features are divided into L groups, and features within each group are considered to have similar properties or belong to the same category.
Step 3: Setting the initial value
Set $l = 1, n = 1, A_n = \emptyset$.
Use the k-nearest neighbor imputation algorithm to initially replace missing values. Sort the variables based on the number of missing values in the original data:$N(z_1) \geq N(z_2) \geq \cdots \geq N(z_p)$, where$N(z_j)$ represents the number of missing cells in the variable z_j.
Step 4: Estimated values of regression coefficients
For a given λ_l=, original missing cells M_l=are interpolated and the estimated values of regression coefficients are obtained from the observed cells O_l in the variable z_l, and replace missing parts $\hat{z}_l^{M_l} = z_{-l}^{O_l}\hat{\beta}^{*AGLASSO}$ using the estimated regression coefficients:

$$\hat{\beta}^{*AGLASSO}(\lambda) = arg\ min\left\{\left\|z_l^{M_l} - \sum_{l=1}^{L} z_{-l}^{O_l}\beta_l\right\|_2^2 + \sum_{l=1}^{L}\lambda_l\hat{\omega}_l|\beta_l|\right\} \tag{8}$$

Step 5: Let $l = l + 1, n = n + 1$, update A_n.
Step 6: We repeat Step (1)-(5) until we have traversed all variables.

Step 7: Repeat Step (4)-(6) until the missing parts are stabilized,
Let $\sum_i(\hat{z}_l^{M_l} - \tilde{z}_l^{M_l})^2 < \delta$ =for all $i \in M_l$=for a small constant δ=where $\hat{z}_l^{M_l}$=is the estimated value of the current iteration and $\tilde{z}_l^{M_l}$ is the i-th estimated value of the previous iteration.
Step 8: Result analysis and comparison
We will present the criteria for evaluating the predictive accuracy of the model in the next subsection. The framework structure of the imputation method for compositional data with missing values is shown in Figure 1.

Pseudocode. The algorithm for AGLASSO imputation method.

```
Input: transformation $ilr(x_i) = z_i = (z_{i1}, z_{i2}, \cdots, z_{ip}), x_i \epsilon S^D, z_i \epsilon R^p, p = 1,2,\cdots,D-1$=, $i = 1,2,\cdots,n$=, features are divided into $L \epsilon \{1,2,\cdots,p\}$ groups.
Initialization: Sort the variables $N(z_1) \geq N(z_2) \geq \cdots \geq N(z_p)$,=set $l = 1, n = 1, A_n = \emptyset, A_n = \{j: \hat{\beta}_j \neq 0\}, j = 1,2,\cdots,p$.
Define $\hat{\beta}^{*AGLASSO}(\lambda) = arg\ min\left\{\left\| z_l^{M_l} - \sum_{l=1}^{L} z_{-l}^{O_l}\beta_l \right\|_2^2 + \sum_{l=1}^{L} \lambda_l \widehat{\omega}_l |\beta_l|\right\}, M_l \subset \{1,2,\cdots,n\}$=is the missing cells, $O_l \subset \{1,2,\cdots,n\} \backslash M_l$ is the observed cells.
for $l = 1$to L do
  $\hat{z}_l^{M_l} = z_{-l}^{O_l}\widehat{\beta}^{*AGLASSO}$
    for $i \epsilon \{1,2,\cdots,n\}$ do
    $\tilde{z}_l^{M_l} \leftarrow \hat{z}_l^{M_l}$
    end for
end for
Let $l = l+1, n = n+1$, update $A_n$.
until $\sum_i(\hat{z}_l^{M_l} - \tilde{z}_l^{M_l})^2 < \delta$ .
Output: $\hat{\beta}_l^{*AGLASSO} = \hat{\beta}_l^{*AGLASSO} / \widehat{\omega}_l$.
```

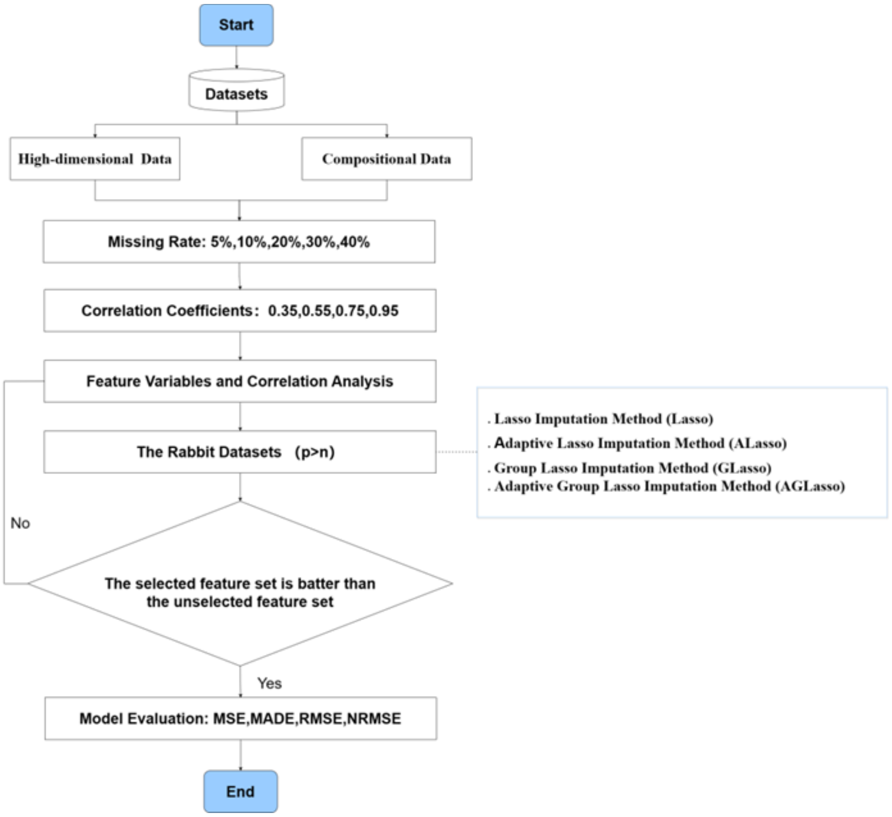
Start
Datasets
High-dimensional Data
Compositional Data
Missing Rate: 5%,10%,20%,30%,40%
Correlation Coefficients: 0.35,0.55,0.75,0.95
Feature Variables and Correlation Analysis
The Rabbit Datasets (p>n)
. Lasso Imputation Method (Lasso)
. Adaptive Lasso Imputation Method (ALasso)
. Group Lasso Imputation Method (GLasso)
. Adaptive Group Lasso Imputation Method (AGLasso)
No
The selected feature set is batter than the unselected feature set
Yes
Model Evaluation: MSE,MADE,RMSE,NRMSE
End

Figure 1. Flow chart for the imputation method for compositional data with missing values

Model Selection Criteria

MSE is a commonly used metric to measure the average squared difference between the actual values and the interpolated values of a model and it defines is given by [39]:

$$MSE = \frac{1}{n_H} \sum_{(i,j)\in H} (\hat{x}_{ij} - x_{ij})^2 \tag{9}$$

where $\hat{x}_{ij}$is the imputed value of x_{ij}, $x_{ij}$is the ture value.$H = \{(i,j): i \in (1,2,\cdots,n); j \in M_i\}$ is the set of column indicators, which represents all rows containing missing values. $M_i \subset \{1,2,\cdots,D\}$ represents the set of missing column indices in x_i. $n_H = |H|$ denotes the number of missing values in the compositional data matrix $X = (x_{ij})_{n\times D}$.

MADE is defined as the mean Aitchison distance between the imputed compositional data $\hat{x}_i$ and the true compositional data x_i:

$$MADE = \frac{1}{n_M} \sum_{i\in M} d_A(\hat{x}_i, x_i) \tag{10}$$

Where $d_A(\hat{x}_i, x_i)$ is the Aitchison distance of the imputed value $\hat{x}_{ij}$ and the ture value x_{ij}. n_M denotes the number rows of missing values in the compositional data matrix $X = (x_{ij})_{n\times D}$.

RMSE is used to assess the accuracy of interpolated models, especially in regression analysis, and it defines is given by [17]:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n_H} \sum_{(i,j)\in H} (\hat{x}_{ij} - x_{ij})^2} \tag{11}$$

NRMSE is a variant of RMSE that provides a normalized measure of prediction error. NRMSE is calculated by [5]:

$$NRMSE = \frac{\sqrt{MSE}}{SD} = \frac{\sqrt{\frac{1}{n_H}\sum_{(i,j)\in H}(\hat{x}_{ij} - x_{ij})^2}}{\frac{1}{n_H - 1}\sum_{(i,j)\in H}(x_{ij} - \bar{x})^2} \tag{12}$$

where $\bar{x} = \frac{1}{n_H}\sum_{(i,j)\in H} x_{ij}$ is the mean value of missing parts in the compositional data matrix $X = (x_{ij})_{n\times D}$. SD is the standard deviation of missing parts in the compositional data matrix $X = (x_{ij})_{n\times D}$.

Experimental Setup

To further validate the effectiveness of the proposed model in the study, we employed the following methods: AGLasso, ALasso, GLasso, and Lasso will be evaluated based on different parameters and experimental setups.

Experimental Tools and Procedures

The experimental tools include operating software such as R and Python. The specific steps in the experiment are as follows:

1. Generate random numbers with certain compositional data characteristics to create a dataset A.
2. Randomly remove some data from the generated complete random numbers to form a missing data set with different missing rates.
3. Apply imputation methods to the compositional data with missing values to obtain a complete dataset B.
4. Compare the differences between dataset A and dataset B, calculate the MSE, MADE, RMSE, NRMSE for both datasets, and analyze the final imputation effectiveness of the methods used.

Experiment Explanation

In order to conduct appropriate data simulation, this thesis simulates standardized data when establishing datasets. The term "data standardization" refers to scaling data proportionally to fall within a specific small range, also known as the exponentiation of statistical data. It mainly involves two aspects: data homogenization and dimensionless processing. Data homogenization addresses issues with different types of data. Directly summing up indicators of different natures cannot accurately reflect the comprehensive results of different forces. It is necessary to first consider changing the nature of inverse indicator data so that all indicators tend to the same trend in influencing the evaluation scheme. Dimensionless processing mainly addresses the comparability of data [2].

The reason for standardizing the data is that different variables often have different units and magnitudes. In normal data analysis, if there is a significant difference in the degree of difference between original data indicators, it is easy to highlight the role of indicators with higher absolute values in comprehensive analysis. Therefore, removing the unit constraints of the data, transforming it into dimensionless pure numerical values, makes it convenient for indicators with different units or magnitudes to be compared and weighted, which is a common measure in data processing. For the convenience of software programming and effect comparison, this paper uniformly adopts simulated data that has been standardized for research purposes.

Under the missing completely at random (MCAR) mechanism, the performance of the proposed method under different missing rate and correlation levels is studied [12][13].

Data Collection

The rabbit dataset is provided (https://www.ebiacuk/ena/browser/view/PRJEB46755), with accession number PRJEB46755. The simulated datasets were analyzed in the supplementary material in reference, it can be downloaded in the supplementary material in reference [26]. The original data contained 89 samples and 3,937 components (microbial genes). The highest correlation is equal to 0.9991, corresponding to gene No.856. The relative abundance of this gene is ranked 201st. The minimum variance of the log-transformed relative abundance of the components is equal to 0.00117, which also corresponds to gene No.856.

In order to simply show the imputation effect of the Adaptive Lasso method, only the 120 genes of rabbit dataset 89×3937 with the largest empirical variance were used for model inference. The sample size of this dataset is $n = 89$, and the number of random variables is D =120. Figure 2 shows a histogram of the Procrustes correlations for all 3937 reference genes of the components (microbial genes), and total log-ratio variance is 0.1601, calculated based on the isometric logarithmic ratio (ilr) of 3937 microbial genes. Among the 3937 genes, this gene ranks 201st in relative abundance.

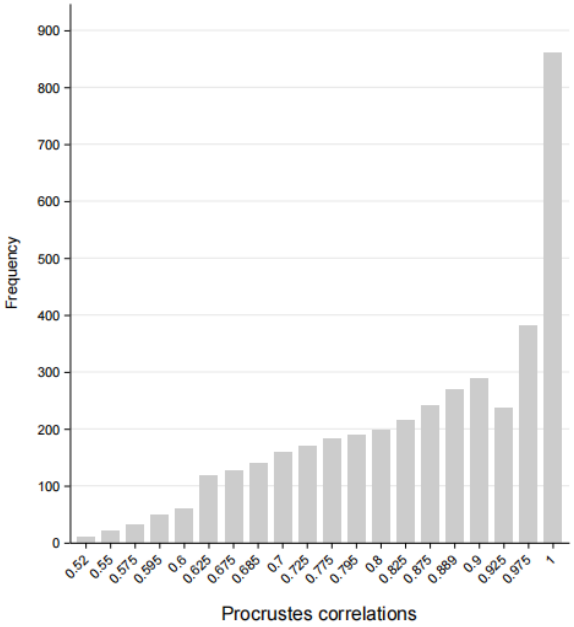

Figure 2. Histograms of Procrustes correlations for each group of all genes in the rabbit dataset calculated using different reference components

The experimental environment was set up on a Windows 10 operating system with 8GB of RAM and a CPU clock speed of 2.4GHz. Table 1 displays the parameters involved in each experiment.

Table 1. Parameters for the proposed imputation method

Parameter Explanation	Parameter Value
Data volume	$n = 89$
Data dimension	$D = 120$
Number of group	$L = 5$
Compositional data in Simplex space	$X = (x_1, x_2, \cdots, x_n)^T$
Compositional data transformed to Real numbers in Euclidean space by isometric logarithmic ratio (ilr)	$Z = ilr(X)$
Number of missing cells in the j-th variable	$N(z_j)$
Coefficients vector of the explanatory variables	$\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$
Imputed value of parameter β	$\hat{\beta}$
Tuning parameter of Lasso	λ
Adjustment parameter of ALasso and AGLasso	λ_n
Tuning parameter of GLasso	$\lambda_l = \lambda \|\hat{\beta}_l\|^{-v}, v > 0$
Adaptive penalty weight of ALasso and AGLasso	$\hat{\omega} = 1/\lvert\hat{\beta}\rvert^{v}, v > 0$

Results and Discussion

First, real data matrix $Z = (z_{ij})_{n\times p}$ is generated from multivariate normal distribution $Z = N^p(\theta, \Sigma_r)$ and then compositional data matrix $X = (x_{ij})_{n\times D}$ =is obtained by inverse transformation of isometric logarithmic transformation ilr^{-1}. To describe the varying degrees of correlation between the components, we set $\theta = (0,0,\cdots,0)_{1\times D}, \Sigma_r = r11^T + (1-r)I,$=where $I^T = (1,1,\cdots,1)_{1\times D}$=, and correlation coefficients $r = 0.35, 0.55, 0.75, 0.95$=and different missing rates 5%, 10%, 20%, 30%, 40%, and set the missing pattern is MCAR (Missing Completely At Random).

The Imputation Results of Different Missing Rates

We use the Lasso imputation methods for the optimal parameter $\lambda = 0.05$, the imputation method based on the ALasso, GLasso and AGLasso for the optimal pair $(v, \lambda_n) = (1,0.5)$. Since $n < D$=, i.e., high-dimensional compositional data with missing values, here we compare AGLasso, ALasso, GLasso, and Lasso imputation effects, and the results of different missing rates are shown in Table 2.

Table 2. Comparison results of different missing rates with all imputation method on the rabbit dataset

MissingRate	Method	Model Evaluation Value			
		MSE	MADE	RMSE	NRMSE
5%	AGLasso	2.36	3.02	0.15	0.10
	ALasso	2.58	3.55	0.37	0.27
	GLasso	3.25	5.32	0.52	0.35
	Lasso	5.10	6.85	0.69	0.51
10%	AGLasso	2.88	3.31	0.21	0.21
	ALasso	3.02	3.69	0.36	0.36
	GLasso	3.54	4.58	0.59	0.58
	Lasso	3.97	5.90	0.94	0.74
20%	AGLasso	4.01	3.87	0.29	0.26
	ALasso	4.68	4.52	0.52	0.65
	GLasso	6.21	6.37	0.71	0.52
	Lasso	7.52	8.09	1.08	0.89
30%	AGLasso	3.54	4.78	0.44	0.45
	ALasso	4.36	5.21	0.68	0.78
	GLasso	5.78	6.80	0.94	0.96
	Lasso	8.54	7.51	1.55	1.80
40%	AGLasso	4.21	4.30	0.67	0.63
	ALasso	6.32	5.28	0.98	0.81
	GLasso	7.25	6.21	1.25	1.52
	Lasso	12.01	9.77	2.54	1.66

Table 2 makes it obvious that the values of the statistical indicators of each variable after filling using the AGLASSO method are the closest to the corresponding results of the complete dataset, with ALASSO and GLASSO being the second most effective and Lasso being the least effective. It can be observed that the fluctuation of deviation from the corresponding results of the complete data set is not significant, in the case of low missing rate (less than 20%); since this paper mainly compares the advantages and disadvantages of the effects of each filling method, the advantages and disadvantages of the filling effects of multiple filling itself will not be discussed.

When the missing rate rises to 20%, the effect of using ALasso is closest to the result of the complete data set. At this point, the results of Lasso are also satisfactory, but they are much worse than the results when the missing rate is 5%. Obviously, as the missing rate increases, this relatively simple and primitive filling method can no longer meet the needs. The effect of AGLasso and GLasso is not much different, and the corresponding statistics of the complete data set are still close to each other. When the missing rate are 30% and 40%, the effect of AGLasso is closest to the standard result, the effect of ALasso and GLasso are second, the drawback of Lasso has been exposed, and the rest of the filling effects are not very satisfactory.

At various missing rates, the RMSE of Lasso imputation is the highest. This is primarily because Lasso imputation does not consider the impact of explanatory variables on the missing data in the regression model when the dependent variable data is missing. It only considers the observed dependent variable and utilizes only the observed values for imputation, resulting in the poorest imputation. On the other hand, at various missing rates, the NRMSE obtained from GLasso imputation and ALasso imputation is very close. That is, the imputed values closely match the true values, and both alternately serve as optimal imputations. The NRMSE obtained from Lasso imputation is slightly worse than that from ALasso imputation and GLasso imputation, but significantly better than Lasso imputations.

The Imputation Results of Different Correlation Coefficients

For different correlation coefficients, we compare all imputation methods to experimental results. Similar as we consider correlation coefficients $r = (0.35, 0.55, 0.75, 0.95)$, when missing rate is 30%. We also use the Lasso imputation methods for the optimal parameter $\lambda = 0.05$, the imputation method based on the ALasso, GLasso and AGLasso for the optimal pair $(v, \lambda_n) = (1,0.5)$. The imputation results of different correlation coefficients are showed in Table 3.

Table 3. Comparison results of different correlation coefficients with all imputation method on the rabbit dataset

Correlation Coefficients	Method	Model Evaluation Value			
		MSE	MADE	RMSE	NRMS E
$r = 0.35$	AGLasso	5.01	5.63	0.64	0.54
	ALasso	5.66	6.79	0.85	0.77
	GLasso	6.52	7.41	1.04	0.86
	Lasso	7.25	8.20	1.28	1.20
$r = 0.55$	AGLasso	5.36	6.21	0.77	0.69
	ALasso	5.90	8.67	0.95	0.82
	GLasso	6.32	9.18	1.48	1.44
	Lasso	8.27	10.25	2.66	2.54
$r = 0.75$	AGLasso	8.21	6.54	0.85	0.79
	ALasso	10.25	8.24	1.02	1.27
	GLasso	12.04	9.57	3.28	3.54
	Lasso	15.18	11.48	6.24	6.57
$r = 0.95$	AGLasso	9.21	7.21	0.84	1.75
	ALasso	10.58	8.94	1.95	3.08
	GLasso	14.76	10.24	3.22	5.21
	Lasso	18.45	19.52	9.84	8.54

As showed in Table 3 the estimator of the AGLasso has preferable performance out-performs the other competitive methods in almost all settings, since the reference components has considered the linear correlation between the variables and is always included in the selected model. The ALasso estimator performs better than GLasso, moreover when the correlation $r = 0.55$, the value of NRMSE based on the ALasso method are slightly smaller than the GLasso and Lasso method, whose error suddenly increases when $r = 0.75$.

Compared with the ALasso and Lasso method, we can also conclude that as the correlation and the dimensionality increase the higher correlation coefficient and dimensionality of variables, the ALasso method imputation effect is better, when $r = 0.95$. This is reasonable because the ALasso method achieve model selection and dimension reduction estimator using penalty function, some regression coefficient directly can down to zero, achieving the purpose of variable selection, at the same time, it can reduce the dimension of data.

When comparing the imputation results of different correlation coefficient component data, the MSE and NRMSE obtained from the AGLasso, ALasso and GLasso imputation methods are significantly superior to those obtained from the Lasso methods. The RMSE values obtained from the GLasso and ALasso imputation methods are also notably better than those from Lasso. Therefore, this study concludes that the sample imputation effectiveness of the AGLasso method is significantly better than that of the other imputation methods.

The results showed that removing certain variables reduced the accuracy of the model. Generally, in the use case of component data, the removed parameters in the model did not improve the accuracy of the model. [22] argued that if the MADE value after removing a parameter is equal to or less than that value, variable selection by grouping improves the interpolation effect of the model, which proves that AGLasso has an advantage over the other models in that it obtains the lowest error with a higher accuracy.

The Imputation Results of Two Patterns

Employ the all imputation method to impute missing values for missing rates and correlation coefficients. Conduct 100 experiments for each missing rate, calculate the value of MSE, MADE, RMSE, NRMSE, and present the results in Figure 3.

As shown in Figure 3, it can be observed that all independent variables selected by the classical ALasso model are significant, indicating a high overall discriminative accuracy for this model. However, in the Lasso models, some independent variables are not significant, yet these models still achieve better discriminative accuracy. This suggests that the ALasso model is effective in selecting variables that better interpolate missing values.

A comparison of the Lasso, ALasso, and AGLasso models reveals that, from theoretical foundations to final performance, the AGLasso model demonstrates superior imputation effectiveness. While the ALasso penalty function is more effective in machine learning under large sample conditions, the AGLasso penalty function, also belonging to machine learning, exhibits better performance when $log\ p > n$. The ALasso function has weaker conditional requirements for harmonic parameters under this circumstance, which aligns with the sample size in this study. Consequently, the AGLasso model outperforms the ALasso model, with a notable difference in the judgment of MSE and RMSE. It is evident that, both being machine learning techniques, the AGLasso penalty function surpasses the Lasso function and is more suitable for research involving component data with missing values, we can draw the following conclusions:

Overall, the ALasso imputed values are significantly smaller than the GLasso and Lasso imputed values and slightly larger than the AGLasso imputed values.

At a missing rate of 20%, the MADE value of AGLasso imputation and the RMSE value of ALasso imputation at a 30% missing rate both fall between the MADE value of Lasso imputation and the RMSE value of ILSR imputation. In the case of a 30% missing rate, AGLasso and ALasso imputation yields the minimum NRMSE value. For a 5% missing rate, the MSE value of Lasso imputation, as well as the MADE value of GLasso imputation at a 40% missing rate, falls between the MSE value of Lasso imputation and the RMSE value of GLasso imputation. This is due to the proximity of the $\hat{\beta}$ values between Lasso imputation, GLasso imputation. For example, when the RMSE values of the three methods are close, the NRMSE value of GLasso imputation is minimized. Therefore, when Lasso imputation and ALasso imputation values are relatively close, GLasso imputation is a novel choice. When GLasso imputation and Lasso imputation are close, but differ significantly from Lasso imputation, GLasso imputation becomes a novel choice.

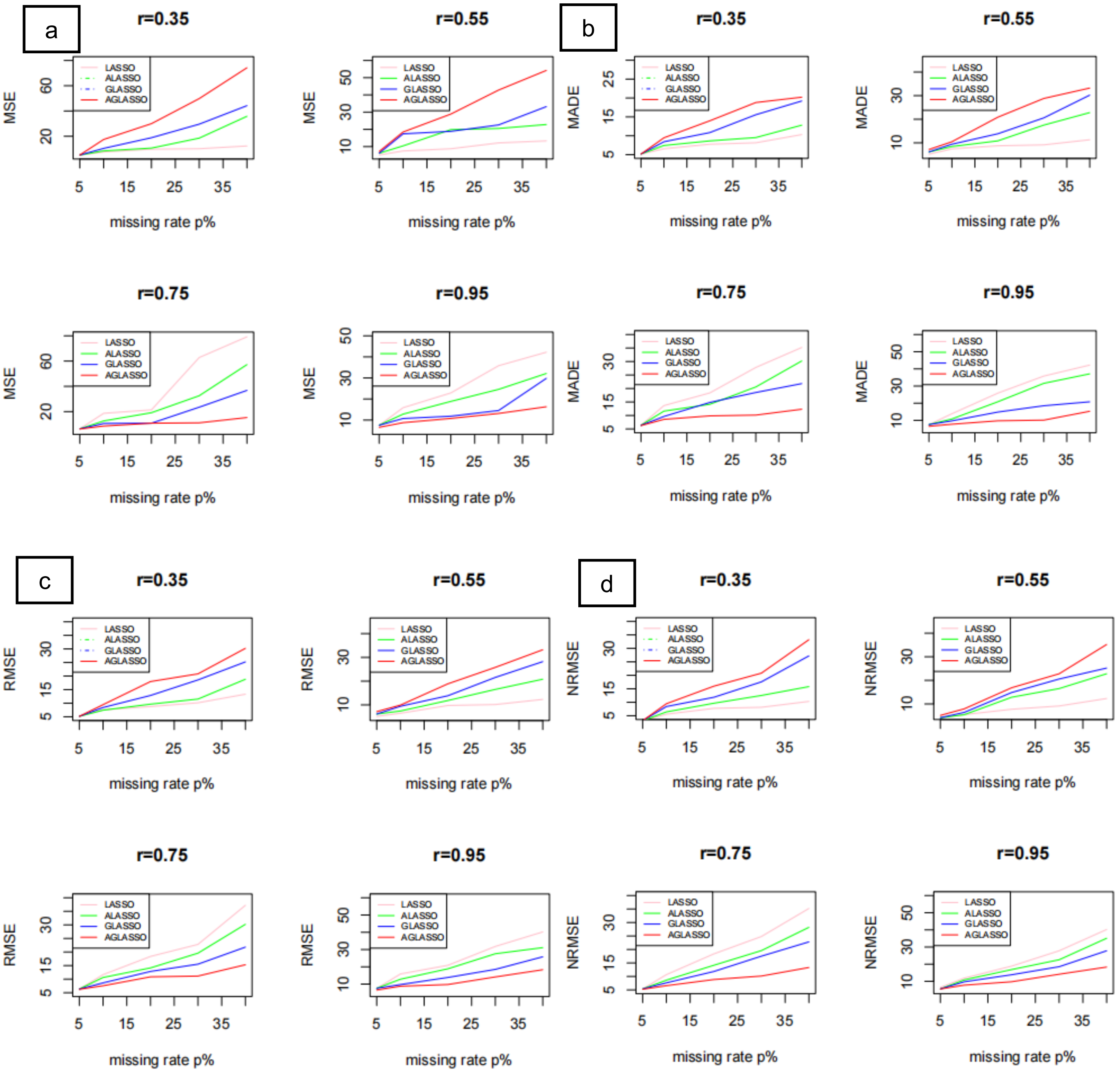

Figure 3. The model selection criteria results of different missing rates and correlation coefficients with all imputation methods

With the increase of missing rate and sample size, the estimated value of the Lasso is farther and farther from the true value, and MSE, MADE, RMSE, NRMSE becomes larger and larger, which shows that the mean imputation method is very poor and only applicable to the case of low missing rate. Lasso is very effective for regression coefficient estimation, but, as the missing rate increases, the estimated values of the scale parameter and the estimated values of the skewness parameter are farther and farther from the true value, and the values of NRMSE gradually increases, and the parameter estimation is poor.

Comparing with GLasso, the parameter estimation effect is significantly improved after AGLasso and ALasso imputation method. The estimation of parameters after modified ALasso imputation method is very good, and the estimation of all parameters is more stable as the missing rate increases. The parameter estimation effect is better than that after ALasso imputation method, and it is the best overall effect of parameter estimation among all imputation methods. Especially, as the missing rate and sample size increases, the above phenomenon is more obvious, which fully illustrates that the AGLasso imputation method is significant effect for the estimation of model parameters after imputation method of missing values.

Conclusions and Discussion

To impute missing values in the imputation model, a variable selection method was employed. Generally, with increasing missing rates, the imputed values from various imputation methods exhibit increased random fluctuations, indicating a deterioration in imputation effectiveness. This is attributed to the reduction in information reflected by the dataset as the missing rate increases, leading to a decrease in available information.

Lasso imputation performed the poorest, AGLasso imputation excelled at high missing rates, and ALasso imputation performed best at low missing rates. GLasso imputation showed slightly larger MSE, MADE, RMSE, and NRMSE values, with ALasso imputation following closely. In scenarios with a certain correlation coefficient for compositional data, as the data's missing rate increases, the imputation effectiveness of Lasso and ALasso methods rapidly decreases, making the estimates not only non-advantageous but even worse.

AGLasso, ALasso, GLasso, and Lasso are all methods used for feature selection and sparse modeling, each with its strengths and weaknesses. In the data analysis context, AGLasso was chosen due to the existence of grouping structures in rabbit genomic data. It allows grouping of features and applies different regularization parameters to each group, better handling feature correlations and multicollinearity. Each group having its own regularization parameter means that features within each group can undergo sparsity or selection to varying extents, enhancing model flexibility. AGLasso aids in more accurately selecting features relevant to the target variable, avoiding issues of over-selection or missing important features.

Unlike traditional Lasso, AGLasso considers the group structure among features, categorizing them into different groups and introducing adaptability with each group having its own regularization parameter. This means that the sparsity or selection extent of features within each group varies, contributing to a better fit to the data's characteristics.

Overall, AGLasso imputation methods significantly outperformed other imputation methods in high-dimensional compositional data, especially in scenarios with high missing data rates and strong correlations. AGLasso imputation demonstrated a clear advantage, yielding more significant imputation results.

In future research on missing data processing will emphasize method innovation, addressing complex data scenarios, and validating practical applications. Interdisciplinary research collaboration and the application of cutting-edge technologies will drive advancements in this field.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

The authors are grateful to the “Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2023/STG06/USM/02/6” for their fnancial support.

References

[1] Ahn, H., Sun, K., & Kim, K. P. (2022). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials & Continua, 70*(1), 767–779.

[2] Aitchison, J., Barcelo-Vidal, C., Martín-Fernandez, J. A., & Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geology, 32*(3), 271–275.

[3] Aşiret, S., & Ömür Sünbül, S. (2023). Effect of missing data on test equating methods under NEAT design. *International Journal of Psychology and Educational Studies, 10*(3), 702–713.

[4] Backdoors, A., Rance, J. C., Zhao, Y., & Mullins, R. D. (2023). Published at ICLR 2023 workshop on backdoor

attacks and defenses in machine learning. *Computer Science, 1–14.*

[5] Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal, 63*(5), 915–947.

[6] Chen, J., & Wu, J. (2023). The prediction of Chongqing's GDP based on the LASSO method and chaotic whale group algorithm–back propagation neural network–ARIMA model. *Scientific Reports, 13*, 15002.

[7] Chen, Q., & Ibrahim, J. G. (2006). Regularization methods for variable selection and estimation in surrogate endpoint regression models with missing data. *Journal of the Royal Statistical Society: Series B, 68*(1), 67–88.

[8] Coenders, G., & Ferrer-Rosell, B. (2020). Compositional data analysis in tourism: Review and future directions. *Tourism Analysis, 25*(1), 153–168.

[9] Combettes, P. L., & Müller, C. L. (2019). Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *Statistics in Biosciences, 13*, 217–242.

[10] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.

[11] Gomer, B., & Yuan, K. (2023). A realistic evaluation of methods for handling missing data when there is a mixture of MCAR, MAR, and MNAR mechanisms in the same dataset. *Multivariate Behavioral Research, 58*(5), 988–1013.

[12] Hammon, A. (2023). Multiple imputation of ordinal missing not at random data. *Advances in Statistical Analysis, 107*, 671–692.

[13] Ji, F., Rabe-Hesketh, S., & Skrondal, A. (2023). Diagnosing and handling common violations of missing at random. *Psychometrika, 88*, 1123–1143.

[14] Ji, H. L., Zhen, T. S., & Zhan, G. (2022). On LASSO for predictive regression. *Journal of Econometrics, 229*(2), 322–349.

[15] Khaliq, A., Sirait, P., & Andri. (2020). KNN imputation missing value for predictor app rating on Google Play using random forest method. *International Journal of Research and Review, 7*(3), 151–160.

[16] Kumar, S., Attri, S. D., & Singh, K. K. (2021). Comparison of Lasso and stepwise regression technique for wheat yield prediction. *Journal of Agrometeorology, 21*(2), 188–192.

[17] Kumar, S., Attri, S. D., & Singh, K. K. (2021). Comparison of Lasso and stepwise regression technique for wheat yield prediction. *Journal of Agrometeorology, 21*(2), 188–192.

[18] Li, C., Pak, D., & Todem, D. (2019). Adaptive Lasso for the Cox regression with interval censored and possibly left truncated data. *Statistical Methods in Medical Research, 29*, 1243–1255.

[19] Li, J., Yan, X., Chaudhary, D., Avula, V., Mudiganti, S., Husby, H. M., Shahjouei, S., Afshar, A., Stewart, W. F., Yeasin, M., Zand, R., & Abedi, V. (2021). Imputation of missing values for electronic health record laboratory data. *NPJ Digital Medicine, 4*, 147.

[20] Liu, Y. (2022). Adaptive Lasso variable selection method for semiparametric spatial autoregressive panel data model with random effects. *Communications in Statistics - Theory and Methods, 70*, 1–19.

[21] Loh, P.-L., & Wainwright, M. J. (2012). High-dimensional regression with missing data: Provable guarantees with non-convexity. *The Annals of Statistics, 40*(3), 1637–1664.

[22] Martínez-Álvaro, M., Greenacre, M., & Agustín Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: A validation of the additive logratio transformation. *Frontiers in Microbiology, 12*, 727398.

[23] Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research, 11*, 2287–2322.

[24] Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(1), 53–71.

[25] Nandhini, S., Debnath, M., Tyagi, S., Mishra, S., & Kumar, K. A. (2020). Stock forecasting using an improved version of adaptive group Lasso. *Journal of Computational and Theoretical Nanoscience, 17*, 3370–3373.

[26] Pandhare, S. C., & Ramanathan, T. V. (2023). The robust desparsified Lasso and the focused information criterion for high-dimensional generalized linear models. *Statistics, 57*(1), 1–25.

[27] Sethi, J. K., & Mittal, M. (2021). An efficient correlation-based adaptive LASSO regression method for air quality index prediction. *Earth Science Informatics, 14*, 1777–1786.

[28] Shen, J., Zhang, Y., & Liu, H. (2020). Combining autoencoders with LASSO for missing data imputation. *IEEE Transactions on Neural Networks and Learning Systems, 31*(5), 1587–1599.

[29] Shi, H., Wang, P., Yang, X., & Yu, H. (2020). An improved mean imputation clustering algorithm for incomplete data. *Neural Processing Letters, 54*, 3537–3550.

[30] Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive Lasso: Variable selection for causal inference. *Biometrics, 73*(4), 1111–1122.

[31] Silva-Ramírez, E., Pino-Mejías, R., & López-Coello, M. (2015). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbors for monotone patterns. *Applied Soft Computing, 29*, 65–74.

[32] Simon, N., & Friedman, J. (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *Journal of Computational and Graphical Statistics, 22*(2), 231–245.

[33] Su, M. H., & Wang, W. J. (2023). A network Lasso model for regression. *Communications in Statistics - Theory and Methods, 52*, 1702–1727.

[34] Tang, J., Zhang, X., Yin, W., Zou, Y., & Wang, Y. (2020). Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory. *Journal of Intelligent Transportation Systems, 25*(5), 439–454.

[35] Tibshirani, R. (2023). High-dimensional regression: Ridge advanced topics in statistical learning, Spring 2023. *Computer Science, Mathematics, 1–14.*

[36] Wang, C., Zhu, R., & Xu, G. (2022). Using Lasso and adaptive Lasso to identify DIF in multidimensional 2PL models. *Multivariate Behavioral Research, 58*, 387–407.

[37] Wang, H., & Leng, C. (2008). A note on adaptive group Lasso. *Computational Statistics & Data Analysis,*

52(12), 5277–5286.

[38] Wang, Y. D., Zhang, W. B., Fan, M. H., Ge, Q., Qiao, B. J., Zuo, X. Y., & Jiang, B. (2022). Regression with adaptive Lasso and correlation-based penalty. *Applied Mathematical Modelling, 105*, 179–196.

[39] Xiao, R., Liu, X., Qiao, H., Zheng, X., Zhang, Y., & Cui, X. (2021). Adaptive LASSO logistic regression based on particle swarm optimization for Alzheimer's disease early diagnosis. *Chemometrics and Intelligent Laboratory Systems, 105*, 104–116.

[40] Xi, L. J., Guo, Z. Y., Yang, X. K., & Ping, Z. G. (2023). Application of LASSO and its extended method in variable selection of regression analysis. *Chinese Journal of Preventive Medicine, 57*, 107–111.

[41] Yuan, H., He, S., & Deng, M. (2019). Compositional data network analysis via Lasso penalized D-trace loss. *Bioinformatics, 35*(18), 3404–3411.

[42] Zhang, C., & Xiang, Y. (2016). On the oracle property of adaptive group Lasso in high-dimensional linear models. *Statistical Papers, 57*, 249–265.

[43] Zhao, P., & Rocha, G. (2018). Adaptive group Lasso for missing data imputation in high-dimensional regression. *Biometrika, 105*(3), 685–699.

[44] Zhou, X., Zhao, P., & Gai, Y. (2022). Imputation-based empirical likelihood inferences for partially nonlinear quantile regression models with missing responses. *Advances in Statistical Analysis, 106*, 705–722.

[45] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.