MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Tian *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 19 (2023) 1082-1098

RESEARCH ARTICLE

Application of Imputation Method for Compositional Data with Missing Values based on Adaptive LASSO Model: The Composition of Employment Industry in Taiyuan, China

Ying Tian[a,b], Majid Khan Majahar Ali[a*], Pei Shan Fam[a], Lili Wu[a], Siti Zulaikha Mohd Jamaludin[a]

[a]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Gelugor, Pulau Pinang, Malaysia; [b]Department of Science, Taiyuan Institute of Technology, 030008, Taiyuan, Shanxi, China

Abstract The tripartite industry classification, which divides all economic activities into three parts, is a classification method to reflect the dynamic process of economic development and the historical trend of the change of resource allocation structure.The fact shows that the proportion of each industry has become an important symbol of the level of national economic development. The proportion of each industry is compositional data,which is a kind of complex multidimensional data used in many fields. All components in the compositional data are non-negative and carry only relative information. In practice, there could be missing values in compositional data. However, general statistical analysis methods cannot be firstly used for compositional data with missing values. The complexity of the missing value of compositional data makes traditional imputation methods no longer suitable. Thus, how to carry out effective statistical inference for compositional data with missing values attracts the attention of many scholars, recently. In this paper, we focus on the imputation problem in compositional data containing missing values, and propose an Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO) imputation method to obtain a complete datasets through variable selection and parameter estimation. Then, the new method is simulated and empirically analyzed, and a comparative study with mean imputation, k-nearest neighbor imputation, and iterative regression imputation is conducted. The results show that the ALASSO imputation method has the highest accuracy for different missing rates, dimensions and correlation coefficients.

Keywords: Missing Values, Compositional Data, Adaptive Lasso, Industry Composition.

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

Received: 16 May 2023
Accepted: 7 Nov. 2023

Introduction

Compositional data appear in different disciplines, for example, geology, biology, economics, etc. For example, in geological literature, many geologists are interested in studying the mineral composition of rock samples [3]. Within biology, the studies of cells usually focus on the relative frequencies instead of the absolute amount [15]. In the course of world economic development in the 20-th century, an obvious feature is the rapid rise of the status of the third industry in the whole national economy. Large-scale and efficient logistics, commodity flow and information flow in the third industry link up a large number of production and consumption, so that material production has been greatly developed [31].

With the deepening of industrialization, highly developed financial, insurance and real estate systems have become important links in improving the overall efficiency of the national economy. City as a country or a region's economic development center, it must be economic and cultural development level is higher space system, its transport, trade, finance, service and other economic center function more developed,

so the proportion of the third industry should be higher than the national average level, some developed urban areas the proportion of the third industry will achieve a quite high level [19]. Taiyuan, the capital of Shanxi Province, is the national famous historical and cultural city, the national garden city, the core city of Taiyuan metropolitan area, and the political, economic, cultural, transportation and international exchange center of Shanxi Province. The development of the third industry is higher than the average level of other towns and cities [32]. From 2004 to 2020, the proportion of the added value of the third industry in the GDP of Taiyuan has increased from 42% to 57.8%. However, Taiyuan as a capital city, its third industry still has a lot of room for development. In order to observe the changes of three industrial structures in Taiyuan's GDP, the charts of industrial structure in 2004, 2011 and 2021 are listed in Table 1.

Table 1. Composition of Taiyuan's there industries

Year	Primary Industry	Secondary Industry	Tertiary Industry
2004	42.0%	55.0%	43.0%
2011	52.8%	45.6%	1.60%
2020	57.8%	41.3%	0.90%

For a long time, Taiyuan has attached great importance to industrial infrastructure and industrial production in the development process, but seriously neglected the development of urban infrastructure and primary and tertiary industries, resulting in urban construction and management lagging behind economic development, low investment ratio in urban infrastructure, too much historical "debt"; the development of agriculture has long been at a low level, urban and rural. Agricultural development has been at a relatively low level for a long time, and the characteristics of the "dualistic" structure of urban and rural areas are obvious; the tertiary industry has been underdeveloped for a long time, and both the scale and structure cannot meet the objective needs of social production and people's life.

From the viewpoint of industrial structure changes in Taiyuan, there are still a lot of problems in the internal structure of Taiyuan industry, and the industrial structure needs to be optimized. Taiyuan should be positioned as a new city, reform some backward, high energy consumption, high pollution enterprises, and develop some new industries, forming a "three, two, one" industrial pattern with heavy industry as the leading industry and secondary and tertiary industries leading economic growth [8].

Background Knowledge

In this paper, it is necessary to define that each component in the compositional data is greater than or equal to zero and less than one. This is a very common situation in practical work, so this modeling method can be applied to most application problems.

Compositional data are quantitative descriptions of relative information and is widely used in various of scientific fields, such as the study of time budgets of groups in psychology, mineral compositions of rocks in geology, and household budget compositions and income elasticities of demand in economics, etc. The concept of compositional data can be traced back to the work in [9], and a D - part simplex is defined as:

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

where the value of c is a random positive constant, usually normalized as 1 or 100. If $x = [x_1, x_2, \cdots, x_D]$ is an element of S^D, it is called a D - part composition. $x = [x_1, x_2, \cdots, x_D]$ denotes a vector in S^D, $z = [z_1, z_2, \cdots, z_p](p = D-1)$ denotes a p - part real valued vector.

To deal with the positive component and the sum-constant contraints in compositional data, we consider the isometric logratio (ilr) transformation is [6]:

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

where the component $x = [x_1, x_2, \cdots, x_D] \in S^D$ is converted into a vector of real numbers with p parts $z = [z_1, z_2, \cdots, z_p] (p = D-1)$.

For a compositional data matrix $X = (x_{ij})_{n \times D}$ with n observations and D parts, the isometric logarithmic ratio transformation of the i - th composition vector $x_i = [x_{i1}, x_{i2}, \cdots, x_{iD}]$ is $ilr(x_i) = z_i = [z_{i1}, z_{i2}, \cdots, z_{ip}] (i = 1,2,\cdots, n; p = D-1)$, and the inverse transformation is as follows:

$$\begin{cases} x_{i1} = \exp\left\{-\sqrt{\frac{p}{D}} z_{i1}\right\} \\ x_{ij} = \exp\left\{\sum_{l=1}^{j-1} \frac{1}{\sqrt{(D-l+1)p}} z_{il} - \sqrt{\frac{D-j}{D-j+1}} z_{ij}\right\} (j = 2,3,\cdots, p) \\ x_{iD} = \exp\left\{\sum_{l=1}^{p} \frac{1}{\sqrt{(D-l+1)p}} z_{il}\right\} \end{cases} \tag{3}$$

We transform the compositional data into a standard Euclidean space using the ilr transformation, and then we can apply standard statistical methods, and the ilr transformation makes the variable $z = ilr(x)$ follows a multivariate normal distribution on R^p, its average vector and covariance matrix respectively μ and $\sum$. Consequently, the original compositional data x follows a normal distribution on a single row space S^D.

To measure the difference between two components, we consider the Aitchison distance, which is a commonly used distance measure applicable to compositional data. It is defined in the ilr space. The ilr transformation is a method for converting compositional data into Euclidean space, avoiding the issue of absolute scale and making the data more suitable for applying Euclidean distance or other methods in Euclidean space [1]. In particular, the Aitchison distance between composition vectors $x = [x_1, x_2, \cdots, x_D]$ and $y = [y_1, y_2, \cdots, y_D] \in S^D$ is defined as follows:

$$\begin{aligned} d_{Aitchison}(x, y) &= \sqrt{\sum_{i=1}^{D} \left\{\ln \frac{x_i}{g(x)} - \ln \frac{y_i}{g(y)}\right\}^2} \\ &= \sqrt{\frac{1}{D} \sum_{i=1}^{p} \sum_{j=i+1}^{D} \left\{\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\right\}^2}. \end{aligned} \tag{4}$$

where $g(x)$ and $g(y)$ are the geometric mean of composition vector x and y, respectively.

Existing Imputation Methods

Now researchers have developed many methods for estimating component data with missing values. Imputation techniques can be classified as univariate imputation, such as MEAN imputation, k -nearest neighbor imputation, EM algorithm; and multivariate imputation such as regression imputation, Markov Chain Monte Carlo (MCMC) algorithm, and so on. More discussions can be found in [18]. In the following, we briefly introduce two of them.

1. MEAN Imputation Method

The mean imputation method is the simplest and easy-to-use replacement method, in which the missing values are first replaced by the mean of the already observed data when preprocessing the data, so that the data can be complete and then the correlations between variables can be analyzed using traditional

data mining methods. However, it is important to pay attention to the type of distribution to which the variables in the data belong, which is not described in detail here [28].

2. knn Imputation Method

The k -nearest neighbor imputation method has been proved to be successful for the multivariate data [5]. The knn method finds the most similar k observation in a combination containing missing values by using a distance metric, and then replaces the missing values with the available information of the selected neighbors. The Aitchison distance is generally used to measure the similarity in knn imputation method.

Suppose a composition contains several cells that contain missing values. First, the knn imputation method searches k -nearest neighbors among the observations with available information of the imputation variable, and determines the similarity according to the information of the units that are not missing. Thus, the k observations can change during successive imputations. Secondly, the median value of the corresponding cell of k nearest neighbor is used to replace the missing part. In general, the unit should be transformed according to the overall size of the part.

The knn imputation method also has some shortcomings [12]. First, the researcher had to determine the best number for the nearest neighbor k . Although the parameter k can be found by simulation by randomly setting the absence of observation units, the computational burden is quite high. Second, in synthetic data, small sample sizes can be problematic when searching for nearest neighbors using available information, as it can lead to different neighbors. Now, though, most practical data sets are of reasonable size. Third, the knn imputation cannot fully account for multivariate relationships between components and can only be considered in search of k -nearest neighbor time. It can be seen that the model-based imputation process can improve the quality of imputation.

The k-nearest neighbor imputation method has been proved to be successful for the multivariate data. The method finds the most similar observation in a combination containing missing values by using a distance metric, and then replaces the missing values with the available information of the selected neighbors. The Aitchison distance is generally used to measure the similarity in imputation method.

Suppose a composition contains several cells that contain missing values. First, the imputation method searches k -nearest neighbors among the observations with available information of the imputation variable, and determines the similarity according to the information of the units that are not missing. Thus, the observations can change during successive imputations. Secondly, the median value of the corresponding cell of nearest neighbor is used to replace the missing part. In general, the unit should be transformed according to the overall size of the part.

The imputation method also has some shortcomings [11]. First, the researcher had to determine the best number for the nearest neighbor. Although the parameter can be found by simulation by randomly setting the absence of observation units, the computational burden is quite high. Second, in synthetic data, small sample sizes can be problematic when searching for nearest neighbors using available information, as it can lead to different neighbors. Now, though, most practical data sets are of reasonable size. Third, the imputation cannot fully account for multivariate relationships between components and can only be considered in search of k -nearest neighbor time. It can be seen that the model-based imputation process can improve the quality of imputation.

3. Iterative Regression Imputation Method

Iterative regression imputation is an Iterative Least Squares Regression (ILSR) techniques to impute missing data [30]. For variables that contain some missing values, the previous observations are considered as covariates to establish an appropriate linear regression model, and to impute missing values. Repeating this the process, one can impute all the missing values. Specifically, suppose that variable y_j is recorded with some missing values and variable vector $z=\left(z_1,z_2,\cdots,z_p\right)$ is recorded completely, so the regression model is established as follows:

$$y_j=\beta_0+\beta_1 z_1+\beta_2 z_2+\cdots+\beta_p z_p+\varepsilon_j$$

where $\beta=\left(\beta_0,\beta_1,\beta_2,\cdots,\beta_p\right)$ is regression coefficient, ε_j is a random noise which follows a normal

distribution $N\left(0,\sigma_j^2 I\right)$. We can estimate the regression coefficients β and σ_j^2 by some methods such as ordinary least square and maximum likelihood. In k -th step of imputation, the new parameters are extracted from the posterior predictive distribution of the missing data, denoted as $\left(\beta_0^{(k)},\beta_1^{(k)},\cdots,\beta_p^{(k)}\right)$ and $\sigma_j^{(k)}$ according to the simulation $\left(\beta_0,\beta_1,\beta_2,\cdots,\beta_p\right)$ and σ_j , then following [18], the missing value can be replaced by the following formula :

$$\beta^{(k)}=\beta_0^{(k)}+\beta_1^{(k)}z_1+\beta_2^{(k)}z_2+\cdots+\beta_p^{(k)}z_p+\tau\sigma_j^{(k)}$$

Where $z=\left(z_1,z_2,\cdots,z_p\right)$ is observed values of the explanatory variables, τ is a normal deviation.

For the iterative regression imputation method, a data set of random residual need to be constructed. There are many methods of construction, for example, we stratify the entire sample according to the explanatory variables z , then in each stratum, put the deviations between the observed value and the average value as a residual. After getting y_j by regression method, in the residual set of the layer, we randomly extract residual item, and take it as the imputation of missing values, that is, $y_j=\hat{y}_j+\varepsilon_j$, this method overcomes the distortion problem of sample distribution.

It is worth noting that the relationship between the variables tends to artificially increase when using rigorous fitted regression equation to predict the target value. Therefore, only when the correlation between explanatory variables and target variables is higher, the regression imputation is more effective [23].

In general, both of the above-mentioned imputation methods have more or less their own advantages and disadvantages. In order to make the imputation effect more stable, more reliable and faster to handle the compositional data with missing values, we propose a novel imputation method -- Adaptive Lasso regression imputation.

Methodology of the Adaptive Lasso Imputation

In this section, we propose a new imputation method innovatively for compositional data with missing values--the Adaptive Lasso imputation. A fast and efficient computational algorithm is also given in this section.

For simplicity, we define the following notations. Let

$$X=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}=(x^{(1)},x^{(2)},\cdots,x^{(D)})=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1D}\\x_{21}&x_{22}&\cdots&x_{2D}\\\vdots&\vdots&\ddots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{nD}\end{pmatrix} \tag{5}$$

be the compositional data matrix, where n is observations, and D is compositions. Using the isometric logarithmic ratio transformation $ilr(x_i)=z_i=(z_{i1},z_{i2},\cdots,z_{ip})(p=D-1)$ the compositional data in monomorphic space are transformed into real numbers in Euclidean space, so that we can use traditional statistical methods to deal with the problem in question [7].

Let the subset $m_l\subset\{1,2,\cdots,n\}$ as the index set of observations that were originally missing in the variable x_l , and $o_l=\{1,2,\cdots,n\}\setminus m_l$ is the index set of observed cells corresponding to x_l . The $z_l^{o_l}$ and $z_l^{m_l}$ denote l - th balance the observed values and missing values, the corresponding variables are x_l . Let the $Z_{-l}^{o_l}$ and $Z_{-l}^{m_l}$ to represent the indicators of missing and non-missing data in x_l , while $Z_{-l}^{o_l}$ and $Z_{-l}^{m_l}$ are the covariates needed during the solving process.

The Adaptive Lasso is an improved version of Lasso regression [27]. The Adaptive Lasso (ALASSO) method uses different weights to punish the coefficients twice. The function of the penalty weights is to make the penalty smaller for the more important variables, so that the important variables can be

selected more easily, while the unimportant variables can be eliminated. This makes up for the defects of Lasso and satisfies the Oracle properties. In addition, the algorithm for solving the Adaptive Lasso problem can be used to solve the Lasso problem with good results [13].

The Lasso estimation is defined as

$$LASSO: \hat{\beta} = \arg_{\beta} \min \left\| y - X\ \beta \right\|^2, subject\ to \left\| \beta \right\|_{l_1} \le t, t \ge 0. \tag{6}$$

Equivalent to

$$LASSO: \hat{\beta} = \arg_{\beta} \min \left\{ \left\| y - X\ \beta \right\|^2 + t \sum_{j=1}^{p} \left| \beta_j \right| \right\}. \tag{7}$$

Similarly, the Adaptive Lasso estimator is defined as

$$ALASSO: \hat{\beta}^{(n)} = \arg_{\beta} \min \left\{ \left\| y - X\ \beta \right\|^2 + t_n \sum_{j=1}^{p} \hat{\omega}_j \left| \beta_j \right| \right\}. \tag{8}$$

where t_n is the adjustment parameter used to balance the penalty term and empirical risk, and the variation of t_n with n ; $\hat{\omega} = 1/\left| \hat{\beta} \right|^{\upsilon}$, $\upsilon > 0$ is the adaptive penalty weight, which works by making the penalty less for the more important variables. The penalty term expression is $t_n \sum_{j=1}^{p} \hat{\omega}_j \left| \beta_j \right|$, 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 [2].

It is easy to see that the main improvement of the adaptive Lasso is the possibility to assign different weights to different coefficients. If we pick a suitable t_n , the adaptive Lasso has the Oracle property, and a larger t_n means a larger penalty to the linear model, which can compress more regression coefficients to 0, then the adaptive lasso estimation satisfies sparsity and asymptotic normality [10,24].

Without loss of generality, we construct $\hat{\beta}(ols)$ the adaptive penalty weight ω . Parameter t_n is used to adjust the sparsity of the model. If the value of t_n is too large, it may lead to substantial deviation in the estimation of large regression coefficients; If the value of t_n is too small, the solution of the model may not be sparse enough. Therefore, the value of t_n should be selected by certain criteria, such as BIC (Bayesian Information criterion), CV (cross validation function), GCV (generalized cross validation function), we use the cross-validation method to determine the optimal t_n and find an optimal pair (υ, t_n) Assume $A_n = \left\{ j : \hat{\beta}_j^{(n)} \ne 0 \right\}, j \in (1,2,3,\cdots,p)$,and the imputed value of parameter β is $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_{o_i})^T$, the prediction vector of the response variable y is $\hat{y}$:

$$\hat{y} = X\ \hat{\beta} = \sum_{j=1}^{p} X_{\ j} \hat{\beta}_j \tag{9}$$

and the sum of squared error is

$$S(\hat{\beta}) = \left\| y - \hat{y} \right\|^2 = \sum_{i=1}^{o_i} (y_i - \hat{y}_i)^2 \tag{10}$$

Denote

$$T(\hat{\beta}) = \sum_{j=1}^{p} \left| \hat{\beta}_j \right| \tag{11}$$

The ALASSO estimator $\hat{\beta}$ is minimizes $S(\hat{\beta})$ subject to a bound of the penalty parameter t_n on $T(\hat{\beta})$. The method of variable selection is essentially the reduction of a complex variable selection problem to a minimization problem of the objective function, the algorithms used to solve the minimization problem are commonly the least angle regression algorithm and the local quadratic approximation algorithm. In this paper, the least angle regression algorithm is used to perform variable selection [4, 25].

The compositional data with missing values, an iterative algorithm via ALASSO can be summarized as follows:

Step (1): Set $l = 1, n = 1, A_n = \varnothing, X_{\ j}^{\ *} = X_{\ j} / \hat{\omega}_j, j = 1,2,\cdots,p$.

Step (2): Here all t_n , the ALASSO problem is solved:

$$ALASSO: \hat{\beta}^{*} = \arg_{\beta}\min\left\{\left\|y - \sum_{j=1}^{p} X_{j}^{*}\hat{\beta}_{j}\right\|^{2} + t_{n}\sum_{j=1}^{p}\left|\beta_{j}\right|\right\}. \quad (12)$$

Step (3): We use the estimates of the regression coefficients $\hat{\beta}_{j}^{*(n)} = \hat{\beta}_{j}^{*} / \hat{\omega}_{j}$, $j = 1,2,\cdots,p$ to replace the missing parts $z_{l}^{m_l}$ by

$$z_{1}^{m_l} = \hat{z}_{1}^{m_l} = Z_{-l}^{m_l}\hat{\beta}\,. \quad (13)$$

Step (4): The values that were originally missing in the cells m_l in variable x_l are updated. Note that also the non-missing cells are updated, but the ratios between them do not change.

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.

After fill the missing parts using the above algorithm, one can get a complete data set. Comparing with the original compositional data, we assess the performance of imputation method using the normalized root mean squared error $nrmse$:

$$nrmse = \sqrt{\frac{mean(X^{true} - X^{imp})^{2}}{Var(X^{true})}} \quad (14)$$

where X^{true} is the original data, X^{imp} is imputed data containing missing parts.

This algorithm ensures that all variables included in the regression model keeping the same correlation degree with the current residual, and thus the algorithm performs much faster than forward selection or step forward process, while avoiding missing some important variables [14, 26].

Simulations Study

In this section, the different missing rates, dimensions and the correlations are analyzed respectively [16]. Firstly, we generate an $n \times p$ real data matrix $Z = (z_{ij})_{n\times p}$ from $N^{p}(\theta, \Sigma_{r})$, and then obtain the compositional data matrix $X = (x_{ij})_{n\times p}$ by using the inverse transformation of isometric logratio transformation ilr^{-1} . In order to give the correlation between the components, we let $\theta = (0,0,\cdots,0)_{1\times D}$, $\Sigma_{r} = r11^{T} + (1-r)I$, where $1^{T} = (1,1,\cdots,1)_{1\times D}$, I is the identity matrix of order D .

1. Scenario 1

Comparison of simulation results of ALASSO imputation method with different missing rates and dimensions [29].

We consider $n = 100,200,300$ different missing rate $mr = (5\%, 10\%, 20\%, 30\%)$. and the repeat 100 simulations for each setting, and take the true value $\beta = (3,1,-2)$. For a given $\upsilon = (0.5,1,2)$, we use the cross-validation method to determine the optimal $t_n = (-0.5,0,0.5)$ and find an optimal pair (υ, t_n) by the ALASSO imputation method. Among them, the imputation results of ALASSO imputation method with sample size of $n = 100$ is shown in Table 2, the imputation results of ALASSO imputation method with sample size of $n = 200$ is shown in Table 3, and the imputation results of ALASSO imputation method with sample size of $n = 300$ is shown in Table 4.

Table 2. Comparison results of different missing rate and parameter pairs using the ALASSO method when sample size n=100

Missing Values	Parameter (υ, t_n)	Imputation Value $(\hat{\beta})$	Imputation Value $(nrmse)$
5%	(0.5,-0.5)	(3.184,2.720,-5.369)	30.4%(1.32)
	(0.5,0)	(2.081,1.832,-2.437)	25.3%(1.27)
	(0.5,0.5)	(2.803,2.195,-1.608)	38.4%(2.68)
	(1,-0.5)	(4.250,0.871,-3.021)	27.6%(3.21)
10%	(1,0)	(1.503,1.253,-1.562)	22.1%(1.45)
	(1,0.5)	(2.214,3.021,-1.207)	19.2%(1.87)
20%	(1,-0.5)	(0.035,0.982,-2.351)	31.9%(3.28)

Missing Values	Parameter (υ, t_n)	Imputation Value $(\hat{\beta})$	Imputation Value $(nrmse)$
30%	(1,0)	(2.551,1.284,-3.517)	24.6%(1.21)
	(1,0.5)	(1.335,3.630,-1.201)	20.5%(1.09)
	(2,-0.5)	(7.256,4.784,-2.327)	26.2%(2.58)
	(2,0)	(2.627,4.851,-1.289)	32.4%(1.61)
	(2,0.5)	(2.256,1.023,-2.144)	28.9%(1.45)

Table 3. Comparison results of different missing rate and parameter pairs using the ALASSO method when sample size n=200

Missing Values	Parameter (υ, t_n)	Imputation Value $(\hat{\beta})$	Imputation Value $(nrmse)$
5%	(0.5,-0.5)	(5.021,0.861,-1.885)	22.7%(2.30)
	(0.5,0)	(2.982,1.778,-1.657)	31.1%(3.21)
	(0.5,0.5)	(2.564,0.961,-4.781)	27.8%(1.85)
10%	(1,-0.5)	(4.237,0.737,-1.327)	25.8%(1.55)
	(1,0)	(2.651,2.841,-2.443)	20.1%(2.28)
	(1,0.5)	(3.351,0.881,-1.652)	29.6%(1.99)
20%	(1,-0.5)	(2.224,0.922,-1.871)	39.2%(3.27)
	(1,0)	(2.630,1.831,-3.681)	32.2%(1.34)
	(1,0.5)	(7.307,0.234,-1.327)	23.7%(2.18)
30%	(2,-0.5)	(2.843,0.627,-1.853)	28.4%(3.48)
	(2,0)	(5.853,3.701,-2.271)	32.9%(2.36)
	(2,0.5)	(2.761,0.257,-1.337)	26.5%(5.02)

Table 4. Comparison results of different missing rate and parameter pairs using the ALASSO method when sample size n=300

Missing Values	Parameter (υ, t_n)	Imputation Value $(\hat{\beta})$	Imputation Value $(nrmse)$
5%	(0.5,-0.5)	(3.241,1.782,-1.531)	42.7%(3.17)
	(0.5,0)	(2.662,0.907,-2.902)	28.1%(2.64)
	(0.5,0.5)	(2.307,0.264,-1.337)	25.4%(1.55)
10%	(1,-0.5)	(1.154,3.322,-4.661)	27.6%(1.34)
	(1,0)	(2.664,0.027,-1.881)	30.2%(2.61)
	(1,0.5)	(2.981,0.634,-0.726)	28.8%(1.08)
20%	(1,-0.5)	(1.017,1.854,-1.461)	33.7%(0.27)
	(1,0)	(2.901,0.227,-1.037)	26.4%(1.36)
	(1,0.5)	(4.087,2.663,-3.291)	27.2%(2.07)
30%	(2,-0.5)	(2.364,0.277,-1.301)	20.1%(1.37)
	(2,0)	(2.940,0.985,-2.385)	32.9%(3.20)
	(2,0.5)	(6.027,1.248,-1.942)	25.7%(1.91)

From Table 2, Table 3 and Table 4, it can be seen the imputation values of the ALASSO method for different sample sizes and missing rates. We find the imputation value $(\hat{\beta})$ and $(nrmse)$ are getting bigger as the sample size increases from $n=100$ to $n=300$ with increasing missing rates, when the adjustment parameter $t_n=0.5$ and the adaptive penalty weight $\upsilon=0.5$. So, the subsequent Adaptive Lasso method will use the optimal parameter pair $(\upsilon, t_n)=(0.5,0.5)$ to impute the missing values.

With the same sample size n, for example, in Table 3, the value of the tuning parameter t_n increases as the missing rate increases, the value of $(nrmse)$ is also increasing, and the interpolation effect of ALASSO is decreasing. This indicates that the value of t_n is too large and the interpretability of the model is not satisfactory.

In the case of the same missing rate, for example, in Table 3, when the missing rate is 10%, as the value of the tuning parameter t_n and the adaptive penalty weight increases, the interpolation effect of ALASSO imputation method will decrease rapidly with the increasing $(nrmse)$, so that the estimated values obtained are not only not advantageous but even worse.

2. Scenario 2

For different miss rates and dimensions, we compare various imputation methods to simulate experimental results [17].

Similar as we consider $n=100,200,300$ different missing rate $mr=(5\%,\ 10\%,\ 20\%,\ 30\%)$.

We use the imputation methods based on penalty function (LASSO), the imputation method based on the Adaptive Lasso for the optimal pair $(\upsilon,t_n)=(0.5,0.5)$ (ALASSO), the k -nearest method based on Aitchison distance (knn), the mean of the observed parts in corresponding component (MENA), and the iterative regression using least-squares estimation (ILSR). Where the imputation results of (ALASSO, LASSO, knn ,MENA, ILSR) method for a sample size of $n=100$ are shown in Table 5, the imputation results of (ALASSO, LASSO, knn , MENA, ILSR) method for a sample size of $n=200$ are shown in Table 6, the imputation results of (ALASSO,LASSO, knn , MENA,ILSR) method for a sample size of $n=300$ are shown in Table 7 based on 100 Monte Carlo experiments.

Table 5. Comparison results of different missing rate and dimensions with various imputation methods when sample size n=100

Missing Values	Parameter (υ,t_n)	Imputation Value $(\hat{\beta})$	Imputation Value $(nrmse)$
5%	ALASSO	(2.981,0.720,-1.084)	30.5%(1.66)
	LASSO	(3.037,1.720,-1.552)	26.7%(5.20)
	knn	(5.734,3.720,-4.468)	26.4%(3.21)
	MEAN	(3.031,0.720,-3.391)	41.8%(1.29)
	ILSR	(2.973,1.720,-0.764)	26.4%(3.07)
10%	ALASSO	(2.337,1.720,-1.846)	27.8%(1.09)
	LASSO	(1.637,0.720,-9.743)	29.9%(1.37)
	knn	(3.982,5.720,-4.881)	36.7%(6.07)
	MEAN	(4.307,1.720,-1.524)	25.4%(2.37)
	ILSR	(7.620,4.720,-6.988)	26.2%(1.37)
20%	ALASSO	(1.361,0.720,-1.772)	25.1%(1.61)
	LASSO	(2.449,3.720,-1.652)	35.6%(2.71)
	knn	(4.561,0.720,-2.794)	27.4%(3.94)
	MEAN	(3.631,4.720,-1.274)	26.6%(2.07)
	ILSR	(3.027,1.720,-5.631)	35.7%(0.91)
30%	ALASSO	(2.329,1.720,-1.027)	33.1%(1.62)
	LASSO	(3.094,0.720,-5.294)	25.0%(1.37)
	knn	(5.671,3.720,-1.360)	21.1%(2.09)
	MEAN	(2.627,3.720,-3.781)	28.5%(3.97)
	ILSR	(3.554,2.720,-1.661)	33.4%(1.07)

Table 6. Comparison results of different missing rate and dimensions with various imputation methods when sample size n=200

Missing Values	Parameter (υ, t_n)	Imputation Value $(\hat{\beta})$	Imputation Value $(nrmse)$
5%	ALASSO	(1.812,2.307,-1.869)	23.0%(3.07)
	LASSO	(2.630,0.951,-6.552)	29.6%(1.02)
	knn	(3.351,4.037,-1.367)	45.5%(2.07)
	MEAN	(0.982,4.942,-0.027)	26.3%(5.37)
	ILSR	(4.037,1.227,-1.981)	32.2%(6.30)
10%	ALASSO	(2.207,0.631,-4.627)	29.6%(2.07)
	LASSO	(2.851,1.027,-1.631)	25.3%(1.37)
	knn	(4.037,3.907,-7.027)	20.2%(1.20)
	MEAN	(1.559,5.861,-1.961)	33.6%(6.07)
	ILSR	(3.840,2.607,-1.207)	25.5%(2.66)
20%	ALASSO	(2.527,1.308,-3.327)	29.3%(2.09)
	LASSO	(1.752,1.255,-1.395)	40.0%(3.91)
	knn	(3.523,6.861,-5.566)	26.1%(1.01)
	MEAN	(4.657,4.665,-1.329)	22.9%(4.05)
	ILSR	(5.500,2.352,-2.782)	26.7%(3.32)
30%	ALASSO	(3.135,4.961,-2.128)	25.4%(1.29)
	LASSO	(2.623,1.127,-1.628)	24.1%(2.28)
	knn	(0.650,2.898,-1.862)	36.3%(1.86)
	MEAN	(1.920,3.038,-6.027)	23.7%(3.21)
	ILSR	(2.537,2.954,-1.965)	31.0%(1.62)

Table 7. Comparison results of different missing rate and dimensions with various imputation methods when sample size n=300

Missing Values	Parameter (υ, t_n)	Imputation Value $(\hat{\beta})$	Imputation Value $(nrmse)$
5%	ALASSO	(1.237,1.373,-2.338)	28.9%(6.30)
	LASSO	(2.636,0.661,-1.524)	30.2%(1.38)
	knn	(7.851,3.162,-3.462)	29.3%(2.39)
	MEAN	(0.207,3.661,-3.620)	25.4%(0.27)
	ILSR	(1.711,4.021,-7.038)	26.6%(3.61)
10%	ALASSO	(2.521,0.631,-0.950)	44.0%(3.20)
	LASSO	(2.880,1.871,-1.630)	29.2%(5.25)
	knn	(4.747,2.651,-4.368)	20.1%(5.90)
	MEAN	(3.961,6.884,-3.983)	18.2%(2.07)
	ILSR	(5.550,1.851,-3.360)	22.1%(3.12)
20%	ALASSO	(4.274,0.651,-1.038)	26.3%(3.38)
	LASSO	(2.895,0.631,-1.446)	25.6%(2.09)
	knn	(0.451,3.884,-1.607)	32.9%(1.21)
	MEAN	(3.337,5.531,-3.492)	23.7%(1.39)
	ILSR	(6.094,2.981,-0.451)	23.0%(3.05)
30%	ALASSO	(1.271,1.521,-2.963)	32.0%(1.33)
	LASSO	(2.961,1.784,-2.850)	22.9%(2.08)
	knn	(4.320,5.964,-1.637)	40.9%(1.99)
	MEAN	(3.027,2.514,-4.960)	23.7%(4.45)
	ILSR	(3.664,3.367,-1.861)	22.3%(6.60)

Observing Tables 5, 6 and 7, we can draw the following conclusions.

(a) With the increase of missing rate and sample size, the estimated value of the mean imputation method(knn) is farther and farther from the true value, and the normalized root mean squared error $(nrmse)$ becomes larger and larger, which shows that the mean imputation method is very poor and only applicable to the case of low missing rate.

(b) Iterative Least Squares Regression (ILSR) 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.

(c) Comparing with iterative regression imputation method, when the sample size increases, the parameter estimation effect is significantly improved after LASSO imputation method.

(d) The estimation of parameters after modified LASSO imputation method (ALASSO) 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 LASSO 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 ALASSO imputation method is significant effect for the estimation of model parameters after imputation method of missing data.

3. Scenario 3

For different miss rates, dimensions and correlation coefficients, we compare various imputation methods to simulate experimental results [20-22].

Similar as we consider $n=100,200,300$ correlation coefficients $cc=(0.35,0.55,0.75,0.95)$ and different missing rate (from 5% to 40% by 5%) respectively. We use the imputation methods based on penalty function (LASSO), the imputation method based on the Adaptive Lasso for the optimal pair $(\upsilon,t_n)=(0.5,0.5)$ (ALASSO), the k - nearest method based on Aitchison distance (knn), the mean of the observed parts in corresponding component (MENA), and the iterative least-squares regression estimation (ILSR). The imputation results results are showed in Figure 1, Figure 2 and Figure 3 based on 100, 200 and 300 Monte Carlo experiments.

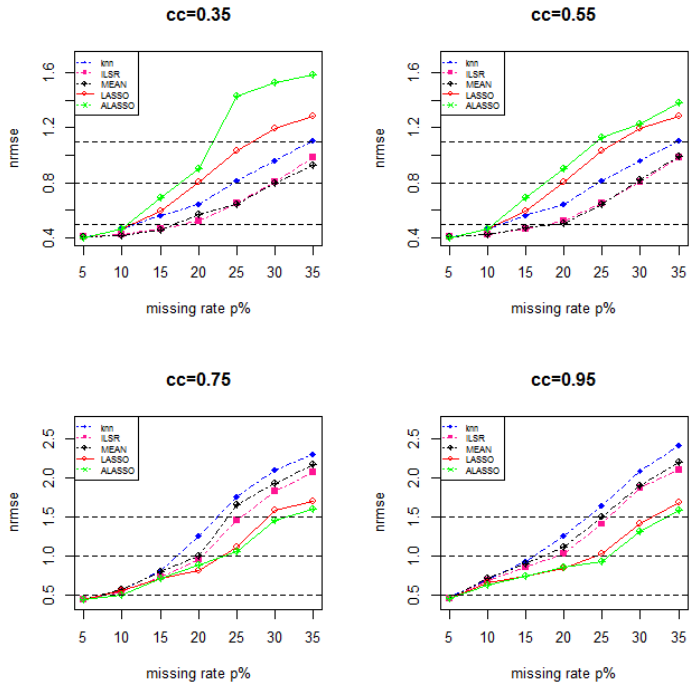

Figure 1. The simulation results of several imputation methods when sample size n=100

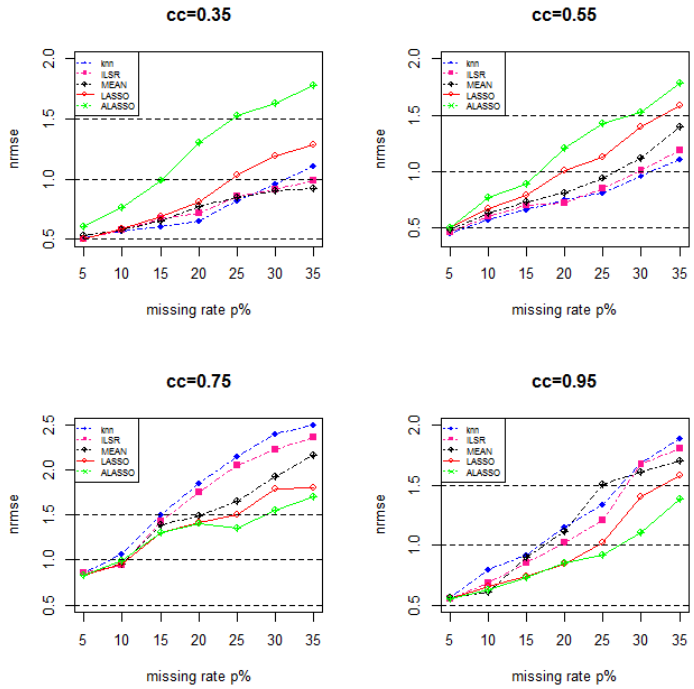

Figure 2. The simulation results of several imputation methods when sample size n=200

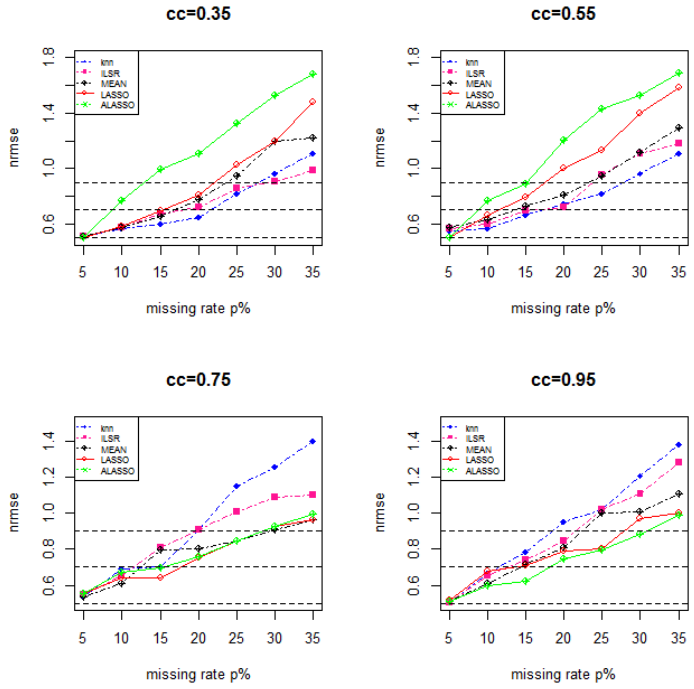

Figure 3. The simulation results of several imputation methods when sample size n=300

As showed in Figure 1, Figure 2 and Figure 3 the estimator of the iterative method using least squares regression estimation (ILSR) 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

ILSR, moreover when the correlation is greater than 0.75 and the missing rate is larger than 15%, the value of $nrmse$ based on the ALASSO method is slightly smaller than the iterative regression method and knn method, whose error suddenly increases when the missing rate is lager than 15%.

Compared with the knn and ILSR method, we can also conclude that as the correlation and the dimensionality increase the higher correlation coefficient and dimensionality of variables, the ALASSO method interpolation effect is better. 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. For compositional data, the ALASSO method also built linear model between variable, made the explanatory of the model better.

Application to Industry Composition of Employed Personnel in Taiyuan

This section will use compositional data for the three industrial composition of employed personnel in Taiyuan, China, from 1991 to 2020. The dataset is provided by the official website of China Statistical Yearbook and contains 30 observations for 20 variables $n=30, D=20$ and the variables

$x=(x_1,x_2,\cdots,x_{20}) = (Agriculture, Forestry, Mining,\cdots, Education, International, Organizations)$

and satisfies the fixed sum limit $\sum_{j=1}^{D} x_{ij} = 100, (i=1,2,\cdots,n)$ .The correlation coefficient matrix for this data set is calculated as follows:

$$\begin{pmatrix} 1.00 & 0.76 & 0.63 & 0.77 & \cdots & 0.82 & 0.53 & 0.57 & 0.89 \\ 0.76 & 1.00 & 0.46 & 0.51 & \cdots & 0.32 & 0.45 & 0.68 & 0.78 \\ 0.63 & 0.46 & 1.00 & 0.36 & \cdots & 0.58 & 0.53 & 0.31 & 0.28 \\ 0.77 & 0.51 & 0.36 & 1.00 & \cdots & 0.69 & 0.72 & 0.61 & 0.34 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0.82 & 0.32 & 0.58 & 0.69 & \cdots & 1.00 & 0.33 & 0.28 & 0.64 \\ 0.53 & 0.45 & 0.53 & 0.72 & \cdots & 0.33 & 1.00 & 0.21 & 0.40 \\ 0.57 & 0.68 & 0.31 & 0.61 & \cdots & 0.21 & 0.21 & 1.00 & 0.25 \\ 0.89 & 0.78 & 0.28 & 0.34 & \cdots & 0.40 & 0.25 & 0.25 & 1.00 \end{pmatrix}$$

The above characteristics of the correlation coefficient matrix indicate that the variables in the Taiyuan industrial structure data are highly correlated with each other, and the missing values can be interpolated by regression equations, and the interpolated values are not too bad.

Let $ilr(x_i) = z_i = (z_{i1}, z_{i2}, \cdots, z_{ip})(i=1,2,3\cdots,30; p=19)$. Since the collected data are complete without missing data, it is assumed for the convenience of the study that the third observation of the sixth variable and the ninth variable are missing, that is X_{36} and X_{39} is considered to be the corresponding variable y_1 and y_2 . Here 5-Fold Cross-Validation(CV) is used to determine the estimates of parameter t_n (see Figure 4). The change process of the variable selection path and activity set of the ALASSO method, the detailed trajectory of the motion is shown in Figure 5.

As can be seen from Figure 4, the Adaptive Lasso imputation method has the smallest CV values for both variables X_{36} and X_{39} at $(\upsilon, t_n) = (0.5, 0.5)$, so we find the optimal parameter pair. As can be seen from Figure 5, the trend of each point in the graph represents the change of the corresponding variable into the model, where the horizontal axis is the total number of steps of the Adaptive Lasso, and the vertical axis is the estimated value of the regression coefficient for each missing value. It is intuitive to see that for the missing values X_{36} and X_{39} , there are nine lines whose values do not converge to 0 at the end, which correspond to the variables, respectively, $(X_1, X_4, X_6, X_8, X_{10}, X_{13}, X_{14}, X_{15}, X_{16})$ and $(X_2, X_4, X_7, X_8, X_9, X_{12}, X_{14}, X_{15}, X_{16})$.

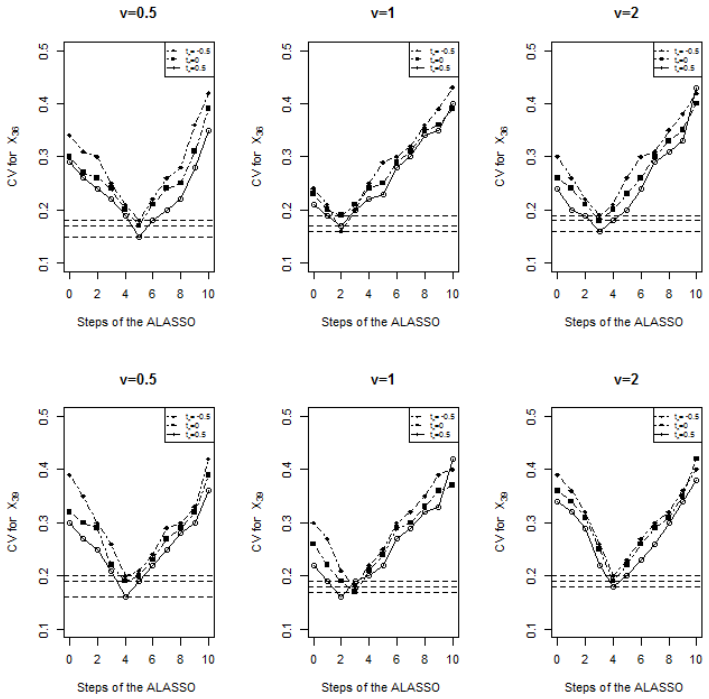

Figure 4. Plots of CV vs. number of steps in the ALASSO for X_{36} and X_{39}

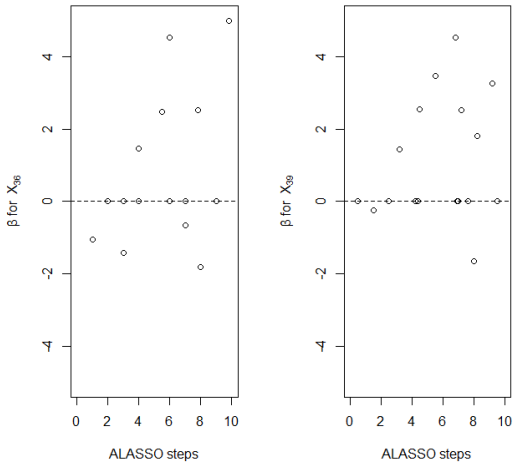

Figure 5. Plots of the ALASSO regression coefficients and the number of steps in the ALASSO for the missing variables

This reflects the fact that the regression coefficients of these variables are not finally compressed to 0, and thus enter the regression model as independent variables to interpolate the missing values in the data. Specifically, the industry composition of employed personnel in Taiyuan data with missing values is selected by ALASSO variables, and the following regression equation can be established:

$$\hat{y}_1 = X_1\hat{\beta}_1 + X_4\hat{\beta}_4 + X_6\hat{\beta}_6 + X_8\hat{\beta}_8 + X_{10}\hat{\beta}_{10} + X_{13}\hat{\beta}_{13} + X_{14}\hat{\beta}_{14} + X_{15}\hat{\beta}_{15} + X_{16}\hat{\beta}_{16} + \varepsilon_1$$

$$\hat{y}_2 = X_2\hat{\beta}_2 + X_4\hat{\beta}_4 + X_7\hat{\beta}_7 + X_8\hat{\beta}_8 + X_9\hat{\beta}_9 + X_{12}\hat{\beta}_{12} + X_{14}\hat{\beta}_{14} + X_{15}\hat{\beta}_{15} + X_{16}\hat{\beta}_{16} + \varepsilon_2$$

where the estimated values of the regression coefficients corresponding to each variable are

$$(\hat{\beta}_1, \hat{\beta}_4, \hat{\beta}_6, \hat{\beta}_8, \hat{\beta}_{10}, \hat{\beta}_{13}, \hat{\beta}_{14}, \hat{\beta}_{15}, \hat{\beta}_{16}) = (-1.05, -1.43, 1.46, 2.47, 4.52, 2.52, -0.65, -1.81, 4.98)$$

$$(\hat{\beta}_2, \hat{\beta}_4, \hat{\beta}_7, \hat{\beta}_8, \hat{\beta}_9, \hat{\beta}_{12}, \hat{\beta}_{14}, \hat{\beta}_{15}, \hat{\beta}_{16}) = (-0.24, 1.43, 2.55, 3.47, 4.52, 2.52, -1.65, 1.81, 3.25)$$

For the optimal pair $(\upsilon, t_n) = (0.5, 0.5)$, we use the ALASSO imputation method introduced in this paper to impute the missing value and compare the results with mean imputation method, k -nearest neighbor method and iterative regression method. The results are summarized in Table 8, these results show that the $nrmse$ of ALASSO is the smallest, indicate that the method proposed in this paper results in a more accurate imputation than other competitive methods in most cases.

Table 8. The results of different optimal pair for the ALASSO method. The values inside the parentheses are the absolute errors between the estimation, and actual observed values respectively is (X_{36} =18.8%) and (X_{39} =2.4%)

Parameter (υ, t_n)	Imputation Value $(\hat{\beta})$	Imputation Value $(nrmse)$
(0.5,-0.5)	15.4%	2.0%
(0.5,0)	15.2%	1.6%
(0.5,0.5)	18.7%	2.3%
(1,-0.5)	16.1%	1.5%
(1,0)	13.2%	1.9%
(1,0.5)	31.5%	6.8%
(2,-0.5)	14.7%	2.2%
(2,0)	20.8%	3.9%
(2,0.5)	12.3%	2.0%

Conclusions and Discussion

In recent decades, the research, theoretical discussion and practical processing of the data have gradually reached a mature stage. A number of good imputation methods have been proposed to deal with missing values in compositional data, and the application areas of each method. In this paper, we systematically discuss the imputation methods of missing values in compositional data and establish the imputation method based on the multiple regression model of ALASSO. The new ALASSO imputation method is compared with LASSO interpolation method, MEAN imputation method, knn interpolation method and ILSR imputation method. The simulation and practical application results show that the ALASSO imputation method outperforms some existing methods in prediction accuracy and variable selection.

The simulations and case analyses of different missing rates, correlation coefficients and dimensionality of composition data containing missing values. The ALASSO method is one of the first methods to achieve both variable selection and parameter estimation, not only to filter variables to make the highly correlated independent variables enter the prediction model, but also to accurately estimate the parameters to be estimated. The core idea is the penalty function, which makes the overall regression coefficients smaller by controlling the reconciliation parameter (υ, t_n), and even makes some regression coefficients tend to 0 or equal to 0.

It is worth noting that the ALASSO interpolation method in this paper has a prerequisite that the linear regression equation is built with the missing values as the dependent variable and the observed data as the independent variables. However, in practical problems, some missing variables are not linearly correlated with other observed variables, which does not satisfy the assumptions of this paper, and further discussion is needed to address this issue.

Currently, LASSO-type methods have been successfully applied in the analysis of large-scale high-dimensional data such as GWAS (Genome-Wide Association Studies) and NGS (Next-Generation Sequencing). For example, Bayesian LASSO has been utilized in a whole-genome association study of Spanish cattle, GLASSO (Group LASSO) has been applied to rheumatoid arthritis whole-genome association research, and the LASSO method has been employed in the analysis of the GAW dataset. With the development and maturation of biochip technology, the analysis of high-dimensional biological data has attracted increasing attention. LASSO-type methods, capable of handling high-dimensional problems, performing variable selection, and parameter estimation, are expected to gain more recognition and usage in the field.

List of Notations

alr: Additive log-ratio transformation.
clr: Centered log-ratio transformation.
ilr: Isometric log-ratio transformation.

m_l : The index set of observation parts in variable x_l .

o_l : The index set of missing parts in variable x_l .

$z_l^{o_l}$: The observed parts of the variable z_l .

$z_l^{m_l}$: The missing parts of the variable z_l .

$Z_{-l}^{o_l}$: The matrices of observed parts in transformed data z .

$Z_{-l}^{m_l}$: The matrices of missing parts in transformed data z .

MCMC: The algorithm of Markov Chain Monte Carlo.

Conflicts of Interest

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

Acknowledgment

The authors wish to thank the China Statistical Yearbook for providing data support. We would also like to express their great appreciation to Dr. Majid Khan Majahar Ali and Fam Pei Shan that assisted by providing necessary information and School of Mathematical Sciences, Universiti Sains Malaysia for the funding.

References

[1] Aitchison, J., Barcelo-Vidal, C., Martın-Fernandez, J. A., Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geology*, *32*(3), 271-275.
[2] Alhusseini, F. H., Flaih, A. N., Alshaybawee, T. (2020). Bayesian extensions on Lasso and adaptive Lasso Tobit regressions. *Periodicals of Engineering and Natural Sciences*, *8*,1131-1140.
[3] Behrouz, R., Fatemeh, A. G., Leila, A., Majid, S., Seyed, H. K. S. (2012). Distribution of metals in sediments of the anzali lagoon, north iran. *Soil and Sediment Contamination: An International Journal*, *21*(6), 768-787.
[4] Courtois, É., Tubert-Bitter, P., Ahmed, I. (2021). New adaptive lasso approaches for variable selection in automated pharmacovigilance signal detection. *BMC Medical Research Methodology*, *21*(1), 271.
[5] Dinh, D., Huynh, V., Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, *571*, 418-442.
[6] Nikaein, H., Sheikhi, A., Gazor, S. (2021). Target detection in passive radar sensors using least angle regression. *IEEE Sensors Journa*l, *21*, 4533-4542.
[7] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology, 35*(3), 279-300.
[8] Feng, Y. (2021). Study on the current situation of economic development and counter measures in the transformation of industrial structure in Shanxi Province. *Economist, 3*, 139-140.
[9] Ferrers, N. M. (1876). *An Elementary treatise on trilinear co-ordinates: The method of reciprocal polars, and the theory of projections.* Macmillan and Company.
[10] Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, *52*(1), 70-84.
[11] Greenacre, M., Martínez-Álvaro, M., Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: A validation of the additive logratio transformation. *Frontiers in Microbiology*, *12*, 1-11.
[12] Hron, K., Templ, M., Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, *54*(12), 3095-3107.
[13] Hui, Z. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429.
[14] Ji, H. L., Zhen, T. S., Zhan, G. (2022). On LASSO for predictive regression. *Journal of Econometrics, 229*(2), 322-349.
[15] Kobayashi, Y. *et al*. (2015). DNA microarray unravels rapid changes in transcriptome of mk-801 treated rat brain. *World Journal of Biological Chemistry*, *6*(4), 389.
[16] 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.

[17] Liu, H., Wang, Y. Y., Chen, W. G. (2020). Three-step imputation of missing values in condition monitoring datasets. *IET Generation, Transmission & Distribution*, *14*(16), 3288-3300.
[18] Little, R. J. A., Rubin, D. B. (2002). Statistical analysis with missing data. John Wiley & Sons. 364-365.
[19] Meng, Y.C. and Li, X. (2020). Analysis of the change of employment industry structure and the effect of capital function in beijing–based on the data of population census (1% population sample survey). *Urban Development Research*, *27*(12), 45-53.
[20] Mostafa, S. M., Eladimy, A. S., Hamad, S., Amano, H. (2020). CBRG: A novel algorithm for handling missing data using Bayesian ridge regression and feature selection based on gain ratio. *IEEE Access*, *8*, 216969-216985.
[21] Muhammadullah, S., Urooj, A., Khan, F.; Alshahrani, M. N., Alqawba, M., Al-Marzouki, S. (2022). Comparison of weighted lag adaptive LASSO with autometrics for covariate selection and forecasting using time-series data. *Complexity. Hindawi*, *2022*, 1-10.
[22] Nijman, S., Leeuwenberg, A. M., Beekers, I., Verkouter, I., Jacobs, J., Bots, M. L., Asselbergs, F. W., Moons, K., Debray, T. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *Journal of Clinical Epidemiology*, *142,* 218-229.
[23] Nordemann, D. J. R., Rigozo, N. R., Echer, E., Souza-Echer, M. P. (2007). Principal components and iterative regression analysis of geophysical series: Application to sunspot number (1750-2004). *Computers and Geosciences*, *34*(11).
[24] 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.
[25] 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.
[26] Shin, D., Lee, S., Jeon, B., Chung, K. (2023). Missing value imputation model based on adversarial autoencoder using spatiotemporal feature extraction. *Intelligent Automation & Soft Computing*, *37*(2), 1925-1940.
[27] Tibshiranit, R. Regression shrinkage and selection via the lasso. (1996). *Journal of the Royal Statistical Society. Series B (Methodological)*, *58(*1), 267-288.
[28] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshiranit, R., Botstein, D., Russ, B. Altman. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525.
[29] 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.
[30] 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.
[31] Han, H., Yu, K. (2022). Partial linear regression of compositional data. *Journal of the Korean Statistical Society*, *51*,1090-1116.
[32] Zheng, S. J. (2015). Optimization of fifinancial structure in shanxi-based on the perspective of industrial specialization. *Economist*, 1, 172-173.