EVALUATING A NEW ADAPTIVE GROUP LASSO IMPUTATION TECHNIQUE FOR HANDLING MISSING VALUES IN COMPOSITIONAL DATA

TIAN YING

UNIVERSITI SAINS MALAYSIA

2024

EVALUATING A NEW ADAPTIVE GROUP LASSO IMPUTATION TECHNIQUE FOR HANDLING MISSING VALUES IN COMPOSITIONAL DATA

by

TIAN YING

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

August 2024

ACKNOWLEDGEMENT

As time swiftly passes and the days and months flow by, my doctoral studies are drawing to a close. I would like to express my heartfelt gratitude to my supervisor, Dr. Majid Khan bin Majahar Ali, and my co-supervisor, Dr. Fam Pei Shan. Reflecting on the three years of studying under Dr. Majid's guidance, my heart is filled with gratitude.

In the initial stages of paper writing, Dr. Majid consistently organized discussions on missing data, laying a robust foundation for subsequent writing. From selecting the research topic, structuring the framework, drafting, to the final revisions, Dr. Majid spared no effort and devoted a significant amount of time. Under his guidance, my capacity to independently research and analyze problems has significantly improved. Studying under Dr. Majid has been the most joyful and meaningful aspect of my student life. His rigorous academic attitude, selfless dedication, and grounded principles have greatly enriched my academic journey.

I sincerely appreciate all the teachers, classmates, and friends who have accompanied me on my journey of learning and growth. Your encouragement, comfort, and companionship along the way have brought me inspiration and enlightenment, providing me with the strength to strive for progress. I extend my gratitude to my husband, Wang Yang, my daughter, Tina, and my parents. Over the past few years, without their understanding, sacrifice, support, and care, completing my research would have been a challenging task. In conclusion, what has passed is merely a prelude, and the future holds promising expectations. Let us move forward together with anticipation.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ALR	Additive-Log-Ratio transformation
CLR	Centered-Log-Ratio transformation
ILR	Isometric-Log-Ratio transformation
GH-EM	The EM algorithm based on Gauss-Hermite integration
MCMC	The algorithm of Markov Chain Monte Carlo
MCAR	The Missing mechanism of Missing Completely At Random
KNN	K-Nearest Neighbors imputation method
MEAN	Mean imputation method
ILSR	Iterative Lest Square Regression
LASSO	Least Absolute Shrinkage and Selection Operator
ALASSO	Adaptive Least Absolute Shrinkage and Selection Operator
AGLASSO	Adaptive Group Least Absolute Shrinkage and Selection Operator
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CV	Cross Validation
GCV	Generalized Cross Validation
MSE	Mean Squaree Error
MADE	Mean Aitchison Distance Error
RMSE	Root Mean Square Error
NRMSE	Normalized Root Mean Square Error

LIST OF APPENDICES

MENILAI TEKNIK IMPUTASI ADAPTIVE GROUP LASSO BAHARU UNTUK MENGENDALIKAN NILAI HILANG DALAM DATA KOMPOSISI

ABSTRAK

Diagram lingkaran adalah bagan statistik yang banyak digunakan untuk merepresentasikan proporsi berbagai komponen dalam entitas tertentu. Bagian data dalam diagram lingkaran, yang juga dikenal sebagai data komposisi, terdiri dari nilai non-negatif, yang hanya berisi informasi relatif. Namun, dalam banyak domain kehidupan nyata, sejumlah besar nilai yang hilang sering kali dikumpulkan. Kompleksitas data komposisi dengan nilai yang hilang membuat metode estimasi tradisional tidak memadai. Dalam tesis ini, sebuah metode imputasi data komposisi yang dirancang berdasarkan LASSO diusulkan untuk menggabungkan metode analisis LASSO kelompok dan LASSO adaptif. Efek estimasi dari data komposisi berdimensi tinggi dan berdimensi rendah dengan nilai yang hilang dibandingkan melalui studi simulasi dan analisis kasus di bawah tingkat kehilangan, dimensi, dan koefisien korelasi yang berbeda. Mempertimbangkan dampak outlier terhadap akurasi estimasi, simulasi dan analisis kasus dilakukan untuk membandingkan algoritma yang diusulkan dengan empat metode yang sudah ada. Hasil eksperimen menunjukkan bahwa metode LASSO kelompok adaptif yang diusulkan menghasilkan kinerja imputasi yang lebih baik, MSE, MADE, RMSE dan NRMSE meningkat hingga 26,6% pada tingkat kehilangan yang dipilih. Penelitian selanjutnya akan menganalisis pengaruh imputasi pada tingkat kehilangan yang terus menerus, mekanisme kehilangan MAR, dan lebih banyak kriteria evaluasi model.

EVALUATING A NEW ADAPTIVE GROUP LASSO IMPUTATION TECHNIQUE FOR HANDLING MISSING VALUES IN COMPOSITIONAL DATA

ABSTRACT

Pie chart is a widely used statistical chart to represent the proportions of various components in a certain entity. The shares of data in a pie chart, also known as compositional data, consist of non-negative values, containing only relative information. However, in many real-life domains, a substantial amount of missing values is often collected. The complexity of compositional data with missing values renders traditional estimation methods inadequate. In this thesis, a compositional data imputation method designed based on LASSO is proposed combining group LASSO and adaptive LASSO analysis methods. The estimation effects of high-dimensional and low-dimensional compositional data with missing values are compared through simulation studies and case analyses under different missing rates, dimensions, and correlation coefficients. Considering the impact of outliers on the accuracy of estimation, both simulation and case analysis are conducted to compare the proposed algorithm against four existing methods. The experimental results demonstrate that the proposed adaptive group LASSO method produces a better imputation performance, MSE, MADE, RMSE and NRMSE increased by up to 26.6% at selected missing rates. Future work analyses the effect of imputation under continuous missing rates, MAR missing mechanism and more model evaluation criteria.

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

With the development and use of cloud computing, applications such as online video, social networking, cloud storage, e-commerce and video surveillance are rapidly emerging (Botha et al., 2022). These Internet and mobile terminal applications have led to a massive increase in the amount of data of all types, and the era of big data has arrived (Martinson et al., 2021; Hashem et al., 2021; Batko et al., 2022) . Finding precise and efficient ways to rationally utilize, effectively process, and efficiently retrieve these big data has become a major issue in many fields.

The application of big data in various fields is becoming more and more widespread, and its analytical and processing power can provide deeper insights and better decision-making support for enterprises, research institutions and governments, bringing innovation and change to many industries (Meen et al., 2023). In healthcare field, big data plays a role in medical research, clinical trials, patient diagnosis and treatment plan optimization. It can help healthcare organizations better manage patient data, improve the efficiency of healthcare services, and even be used for disease prediction and prevention (Ramyanjali et al., 2023) . In financial services, banks and financial institutions use big data for credit assessment, risk management, fraud detection, and stock market analysis (Madem et al., 2023; Zhang et al., 2023). Big data technology supports high-frequency trading, personalized investment advice and customer service.

The research scope of big data processing is very wide, and high-dimensional data processing has always been a research hotspot in data mining (Xia et al., 2023;

Ma et al., 2023) . With the increase of data volume, it is imperative to analyze the high-dimensional data in big data. Due to the sparsity of high-dimensional data, the data processing methods in high-dimensional space are significantly different from those in low-dimensional space. Many mature algorithms in the low-dimensional space cannot achieve the expected results in the high-dimensional space, or even cannot run (Absalom et al., 2020; Sarstedt et al., 2021; Agarwal et al., 2020; Wang et al., 2021). This requires us to improve existing algorithms or propose new frameworks to adapt to the context of big data applications.

In the process of high-dimensional data mining, close to 40% of the University of California Irvine (UCI) datasets (https://archive.ics.uci.edu/) (a benchmark database in machine learning) contains missing values (Mohammed et al., 2017; Untoro et al., 2020) . According to statistics, target identification takes about 20% of the time in the data mining process, data preprocessing takes nearly 60% of the time, while data mining and knowledge analysis takes only 10% of the time. It can be seen that people put more than half of the energy on data preprocessing, which is because most of the real data have problems such as data inconsistency, too many outliers, data redundancy, missing data and so on.

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 (Greenacre et al., 2021; Tian et al., 2018; Javanbakht et al., 2022). 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 (Coenders et al., 2020; Chao

et al., 2022; Ye et al., 2022). For the problem of interpolation of missing values in high dimensional component data, common approaches include, mean or median imputation: for each missing component, interpolation 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 of the missing component and then interpolates using the component values of those samples (Daldiri et al., 2023; As'ad et al., 2023). 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 (Liu et al., 2023; Ahn et al., 2022; Little et al., 2022). 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.

1.2 Problem Statement

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 (Absalom et al., 2020; Sarstedt et al., 2021; Agarwal et al., 2020; Wang et al., 2021). The model obtained by clustering is too sensitive to the clustering algorithm. The method of partial least

squares and principal component regression usually select the components according to the criteria of cumulative contribution rate, the size of the characteristic root, and statistical significance, etc., and the resulting model is biased although the structure is simple and the estimation is stable. Although the obtained principal components may have certain practical significance, they cannot explain the effect of individual covariates clearly. The ridge regression is able to deal with multicollinearity between variables in a better manner, but the results of ridge regression and tree-based integration methods are not satisfactory. Although ridge regression can better deal with multicollinearity among variables, it cannot provide a sparse model because it cannot reduce the dimensionality. The tree-based integration method tends to have poor interpretability due to too many adjustment parameters. LASSO simplifies the model by introducing L1 regularisation for downscaling and variable selection, automatically reducing unimportant coefficients to zero (Fidalgo et al., 2023; Malti et al., 2023; Balcilar et al., 2022). Adaptive LASSO improves on LASSO by adjusting the penalty strength using data-driven weights for more accurate variable selection and parameter estimation (Kumar et al., 2023; Balakrishnan et al., 2023; Feng et al., 2022). Adaptive LASSO is consistent and can more reliably select the correct variables in large samples. In contrast, LASSO has limitations in variable selection stability.

The problem is that LASSO and adaptive LASSO generally use the cross-validation (CV) criterion, because it enables efficient selection of regularisation parameters and adaptive weights, thus balancing model fit and complexity. Cross-validation prevents overfitting and improves the generalisation ability and reliability of the model by dividing the dataset into training and validation sets and evaluating the model's performance on unseen data several times (Zhao et al., 2023; Baldé et al.,

2022; Muhammadullah et al., 2022). However, for small samples or highly noisy datasets, the results of cross-validation may be unstable, leading to large fluctuations in the values of the selected parameter values. When dealing with highly correlated variables, one of the variables may be randomly selected while ignoring other correlated variables, leading to unstable variable selection. These shortcomings can be effectively mitigated by using improved methods such as generalised cross-validation (Maharani & Saputro, 2021; Liu et al., 2022), Akaike Information Criterion (AIC) (Bozdogan, 1987; Arnold, 2010) and Bayesian Information Criterion (BIC) (Lee & Chen, 2020; O'Neill & Burke, 2023) to improve the performance and stability of the model.

Another problem is that missing values are a common challenge in data analysis and can affect the accuracy of statistical analyses and machine learning models. Commonly used missing value interpolation methods in machine learning and deep learning include mean/median filling, constant filling, regression interpolation, KNN interpolation, and multiple interpolation method (MICE). Mean filling is simple and fast but ignores the relationship between features; constant filling is suitable for class features but may introduce bias; regression interpolation is highly accurate but relies on the predictive power of other features; KNN interpolation takes into account the similarity between features but has a high computational overhead; and MICE is suitable for complex models but requires multiple iterations. In deep learning, neural networks can learn and process missing values end-to-end, but require large amounts of data and computational resources. The choice of interpolation method should balance accuracy and computational efficiency based on data characteristics, missing value patterns, and task requirements (Pham et al., 2022; Tashiro et al., 2021; Donlen, 2022; Wolbers et al.,

2021; Ma et al., 2021; Habibi et al., 2018; Lee et al., 2021; Beesley et al., 2021). The main model studied in this thesis is the models used for comparison are the Adaptive Group LASSO model, which can find applications in scenarios where variables are naturally clustered into groups (e.g., gene expression data in biology, features from different sensors in IoT applications), leading to more efficient feature selection and improved model generalisation (Anas et al., 2022; Nugroho et al., 2023).

The last problem is a binary classification dataset containing missing values. Commonly used binary classification methods include logistic regression, support vector machines, decision trees and random forests, neural networks, and plain Bayes (Ganesh et al., 2022; Xu et al., 2023; Naveen et al., 2023; Wu et al., 2022; Wei et al., 2023; Jin et al., 2022). Each of them shows advantages in dealing with nonlinear relationships, high-dimensional data, and complex models, but they also have drawbacks such as overfitting, the need for parameter tuning, and sensitivity to large-scale data and noise. The selection of a suitable method requires comprehensive consideration of data characteristics, model complexity and practical needs. In this thesis, we propose adaptive LASSO-Logistic models that combine adaptive LASSO regularisation techniques with logistic regression for scenarios where feature selection is crucial for building interpretable models (Mahajan et al., 2023; Arora et al., 2023; Chen et al., 2022),such as medical diagnosis, credit scoring, or any binary classification problem with high-dimensional data (Schober et al., 2021; Girish et al., 2023; Reddy et al., 2023).

1.3 Objectives of the Study

The objectives of the study are:

i. To improve the accuracy of regression coefficient valuation in LASSO and Adaptive LASSO imputation method adding the parameter selection criterion of AIC, BIC, GCV.

ii. To increase the error of the imputation method for low-dimensional compositional data using the Adaptive Group LASSO (AGLASSO) imputation method.

iii. To enhance the error of the imputation method for high-dimensional binary categorical compositional data, the Adaptive LASSO imputation method based on the Logistic model (ALASSO-Logit) is used.

1.4 Scope and Limitation

This thesis concentrates on the problem of handling missing values in high-dimensional compositional data. When there are missing values in the high-dimensional compositional data, it is not possible to perform closure operations on the vectors to obtain the corresponding ratios. In addition, the fixed sum of the vector of the high-dimensional compositional data is not known, and traditional statistical methods usually cannot produce reasonable results (Li et al., 2022; Liang et al., 2022; He et al., 2020; Tu et al., 2023; Chen et al., 2018; Engle et al., 2023). Thus, filling in the missing values of the high-dimensional compositional data is an essential prerequisite for compositional data analysis.

Doing deletion process for missing data may discard the hidden value in the data and cause data waste. Usually, we use filling algorithms for missing data. Already existed missing data filling algorithms are: mean filling, median filling, K-Nearest Neighbor (KNN), Singular Value Decomposition (SVD), Bayesian Principal

Component Analysis (BPCA), Minimum Half Filling (MHF), and Minimum Half Filling (MHF), Halfminimum (HM), Multiple Filling, Replicate peak (REP), etc. Among these methods, some of them are too simple and do not fully consider the existing information and the connection between the information items, thus lacking in estimation accuracy, while some of them require a lot of specific preconditioning assumptions, which reduces the scope of application of the methods. Therefore, improving the accuracy of filling while appropriately reducing the preconditions is a key element in data filling.

This thesis proposes a new imputation method - based on the LASSO method: the Adaptive LASSO imputation method, the Adaptive LASSO imputation method based on the Logistic model, the Adaptive Group LASSO imputation method. (Raudhatunnisa et al., 2022; Fadlil et al., 2022; Shi et al., 2020; Anas et al., 2022; Oktaviani et al., 2020).

There are some limitations of this study, for instance, handling missing data typically requires assumptions about the missing mechanisms, such as Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing not at Random (MNAR). The LASSO class methods employed in this thesis are analyzed and evaluated exclusively for MCAR (Waterbury, 2019). The LASSO class methods are all a regularization method for linear regression, and thus are themselves imputation techniques for modeling linear relationships. For nonlinear data, the use of ordinary linear models (including LASSO class) may not be sufficient to capture nonlinear relationships in the data. However, there are ways to extend LASSO to handle nonlinear data: Elastic Networks, Kernel Functions, and so on. Therefore, if the data clearly exhibits nonlinear relationships, it is recommended to use models more suitable for dealing with nonlinear relationships higher dimensional data in

higher dimensional spaces, for example, commonly used nonlinear dimensionality reduction methods such as Isomap, t-SNE, Kernel PCA, autoencoder, etc (Anowar et al., 2021； Han et al., 2022; Uriot et al., 2022). LASSO performs well in dealing with high-dimensional linear relationship problems, but for nonlinear data, it is still important to choose the appropriate imputation method according to the nature of the data and the background of the problem. In the imputation of missing values, only MSE, RMSE and MAE models were used to evaluate the metrics, which are sensitive to outliers. The specificity of the compositional data requires consideration of their scale characteristics and closure properties, hence the introduction of the Aitchison distance (MADE). The goal of the estimation is to recover missing values as accurately as possible, rather than explaining the overall variance, which makes R^2 limited and inappropriate. However, we can attempt to use these methods in future work. Since real data contains both continuous and discrete data, accuracy, F1 scores, AUC and ROC are mainly used in classification tasks to evaluate the performance of classification models, and thus are not applicable to deal with continuous data. However, we can try to use these metrics in the ALASSO-Logit.

1.5 Significance of the study

High-dimensional compositional data are widely available in biomedical, environmental science, finance and other fields, and missing data is a common problem. Effective imputation methods can improve data integrity and analysis accuracy, and avoid bias and information loss caused by missing values. To address the complexity and characteristics of high-dimensional data, developing new imputation methods not only improves model prediction performance, but also reveals underlying data structures and patterns. Especially in machine learning and

statistical analysis, reliable imputation methods can enhance the robustness and generalisation of algorithms, thus improving the quality and credibility of overall data analysis.

High-dimensional compositional data usually contains a large number of features, and direct analysis can lead to high computational complexity, model overfitting, and difficulties in imputation (Emmanue et al., 2021； Lee et al., 2021; Carpenter & Smuk, 2021; van Ginkel et al., 2020; Khan & Hoque, 2020; Muzellec et al., 2021; Chai et al., 2020; Felix et al., 2023; Tawakuli et al., 2023; Kumar et al., 2021). The proposed LASSO and Adaptive LASSO, as powerful dimensionality reduction tools, are able to provide significant advantages in feature selection and dimensionality reduction through the introduction of regularisation. LASSO and Adaptive LASSO methods achieve feature selection and dimensionality reduction through sparsity constraints. In high-dimensional data, many features may be redundant or noisy, and LASSO effectively reduces the dimensionality of the model by applying penalties to unimportant features and reducing their coefficients to zero. This not only reduces the computational complexity but also improves the interpretability of the model.LASSO and adaptive LASSO methods help to improve the generalisation ability of the model. By reducing the number of features, the models are better able to capture the key structure of the data while avoiding overfitting, improving the accuracy and stability of the predictions.

Regularised parameters and adaptive weights in LASSO and adaptive LASSO models control feature selection and model sparsity. Cross-Validation (CV) optimises the model performance by training the model on the training set and evaluating the model performance on the validation set to select the best parameter values. CV although widely used, has some drawbacks. Combining the advantages of

CV, GCV, AIC and BIC can effectively improve the reliability and efficiency of model selection. CV provides direct model validation, GCV provides efficient error estimation, while AIC and BIC help balance model complexity and performance. This multi-criteria combination approach allows for a more comprehensive assessment of the model, optimises model selection and enhances the generalisation ability of the model.

The proposed Adaptive Group LASSO (AGLASSO) imputation method combines the sparsity of LASSO and the group feature selection capability of group LASSO, which is able to efficiently handle high-dimensional data with group structure (Hastie et al., 2020； McEligot et al., 2020; Cao et al., 2021; Setty & Thaler, 2023; Zhao et al., 2023; Chen & Wu, 2023; Xu, 2023; Tibshirani, 2023; Feng et al., 2022; Guo & Wang, 2022). By assigning different regularisation weights to different feature groups, this method is able to select important features more accurately while maintaining the structural information of the data. It not only improves the accuracy of imputation, but also enhances the ability to capture the underlying structure in high-dimensional data. Especially in fields such as biomedicine and finance, where data usually have group features, such as genomic group structure for genomic data or industry group structure for financial data, adaptive group LASSO is able to better handle these complex data. In conclusion, this approach helps to improve the reliability of data analysis and the generalisation ability of models, and promotes the development of scientific research and practical applications.

The proposed Adaptive LASSO Logistic (ALASSO-Logit) regression imputation method can effectively use the existing complete data to predict and fill in missing values by modelling the occurrence mechanism of missing data (Becht et al.,

2018; Jia et al., 2022; Hasan & Abdulazeez, 2021; Cantini et al., 2021; Tripathy et al., 2021; Smithson & Broomell, 2022). This method not only maintains the integrity of the data, but also reduces the loss of information due to missing data and improves the accuracy and reliability of data analysis. In addition, logistic regression imputation is suitable for dichotomous or multiclassified data, and can handle complex missing patterns and nonlinear relationships, making the imputation results more consistent with the actual data distribution. Finally, the method has strong interpretability and computational efficiency, which is suitable for the processing of large-scale data sets.

Missing value imputation methods enrich the theoretical body of data processing and statistical analysis. By exploring the nature and performance of different imputation methods, it is possible to gain a deeper understanding of the impact of missing data on model estimation and to optimise data integrity, e.g. through innovations in algorithmic complexity, robustness and adaptability. In practice, the problem of missing data is prevalent and can lead to information loss and analysis bias. Effective imputation methods can fill in the missing values and restore the integrity of data, thus improving the accuracy of analysis results and the effectiveness of decision-making. Especially in the fields of healthcare, finance and social sciences, missing value imputation can help build more reliable data models to support data-driven decision-making in scientific research and practical applications.

1.6 Thesis Framework

The thesis is organised as follows:

Chapter 1 covers the background to the problem, the problem statement, the objectives of the study, the scope and limitations, and the significance of the study.

Chapter 2 discusses the literature review on dimension reduction methods for high dimensional data, solutions to the problem of fixing and summing of compositional data, imputation techniques for missing values, LASSO methods and the current state of research on outliers.

Chapter 3 is methodology; it provides the flowchart of the study, data collection and the models used in the study are presented. It mainly introduces the definition, properties, transformations and other related knowledge of compositional data; high-dimensional data and its variable selection algorithms; commonly used missing value processing methods in compositional data, and elaborates the definition and specific steps of various algorithms.

Chapter 4 is results and discussion; it is a comparative study of the missing values imputation methods in high-dimensional compositional data. Under a series of different missing rates, dimensions and correlation coefficients, we simulate and empirically analyze the new LASSO method, and then conduct a comparative study with mean imputation method, k-nearest neighbor imputation method, and iterative regression imputation method. It provides the target results accordingly.

Chapter 5 includes conclusions, contributions of the study, limitations of the study and future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In standard multivariate statistical analysis, the problem of missing values is a difficult one. The problem of missing values in compositional data becomes more complicated. The number of dimensions and missing rate of data in various survey results are on the rise, the traditional method of handling missing values is not effective, and the issue of a more effective way to deal with missing data is of considerable concern. The issue of compositional data in Section 2.2. The literature review on high-dimensional data in Section 2.3. Section 2.4 illustrates the literature review on missing data. Section 2.5 reviews the literature on outliers. Section 2.6 illustrates the literature review on machine learning for dealing with big data. Finally, in Section 2.7, the summary of the literature review is presented.

2.2 Compositional Data

Compositional data is data that describes the proportional relationships between parts within a whole (Smithson & Broomell, 2022). It typically appears in the form of ratios or percentages. Researching compositional data mainly involves studying the relative information about the proportions that absolute data occupies within the whole (Greenacre, 2022).

2.2.1 Transformation of Compositional Data

Compositional data cannot be processed directly by traditional statistical methods due to their definitions and limitations. A novel type of transformation,

Reciprocal Logarithmic Ratio (RLR) transformation, is proposed that enables convergent cross-mapping to be applied to compositional data and is expected to provide a better understanding of ecological interactions by estimating causal relationships in compositional data (Kumakura et al., 2021).

Applying cluster analysis to geochemical data, the study compares the effects of various data transformation methods on cluster analysis of geochemical compositional data (Zhou et al., 2018) . The use of the Box-Cox transform as an alternative to the logarithmic transform leads to a significant improvement in existing methods for processing compositional data and partially demonstrates that one of the most fundamental problems that Aitchison managed to overcome can be solved with the use of the logarithmic ratio transform (Rayens & Srinivasan, 1991). Dealing with the field of high-dimensional compositional data, the chiPower transform is proposed to be as close as possible to the logarithmic scale transform without replacing the zeros, an alternative that can present a high degree of consistency and isometric properties (Greenacre, 2022).

Partial linear regression modeling by means of an isometric log-ratio (ilr) transformation provides conditions for the identification of the linear parameters and the development of an estimator based on the expectations of the response and covariates (Han & Yu, 2022) . The model architecture used in this study is the convolutional neural network (CNN) , transfer learning from a VGG-16 model pre-trained on the ImageNet dataset, and a multi-target output layer with softmax regression trained to minimise root mean squared error (RMSE) loss. Adaptive modeling using compositional data to predict biomass and improve prediction errors (Narayanan et al., 2021).

A new class of transforms called isometric α-transforms (α-IT), addresses the problem of proposing models and methods to analyze and predict, through kriging, this type of data (Clarotto et al., 2021). This study uses a ribosomal RNA microbiome dataset and compares the performance of 11 normalization methods and a centered logarithmic ratio (CLR)-based transformation method (Bars-Cortina, 2022). Reanalysis of NGS amplicon and metabolomics component data by analyzing CLR transformations revealed new relationships and stronger associations between sample conditions and microbial and metabolic community profiles (Sisk-Hackworth et al., 2020).

A new method for imputing rounded zeros based on artificial neural networks (ANNs) is presented and compared with conventional methods. Additionally, the log-ratio transformations within the artificial neural network imputation procedure contribute to improved results (Templ et al., 2020). Logarithmic contrasts are introduced as linear functions with respect to parsimonious operations for compositional data problems, using the relationship between internal perturbations and idempotent operations and the usual operations in the space of real numbers (Martín-Fernández, 2019).

Table 2.1 shows the literature about transformation of compositional data. From Table 2.1, the gap that found from the literature is transformations on compositional data are mainly focused on low-dimensional data $n > p$, with most of the study data being small and medium data and very few studies on large data.

Table 2.1 Literature about transformation of compositional data

Author(s) Year(s)	Sample size			Field of data set				Simulations		Data Dimension		Method
	Small $n < 50$	**Medium** $50 \leq n < 100$	**Large** $n \geq 100$	**Medicine**	**Geology**	**Biologic**	**Other**	**Monte Carlo**	**Other**	$p < n$	$p > n$	
(Kumakura et al., 2021)	√			√				√		√		ILR
(Zhou et al., 2018)			√		√			√			√	ILR,CLR,ALR
(Backdoors et al., 2023)	√						√		√	√		Box-Cox
(Greenacre, 2022)			√				√		√		√	chiPower
(Han & Yu, 2022)		√					√	√		√		ILR
(Narayanan et al., 2021)	√			√				√		√		ILR,CNNs
(Martín-Fernández, 2019).	√						√	√		√		Log-Ratio

Table 2.1 (Continued)

Author(s) Year(s)	Sample size			Field of data set				Simulations		Data Dimension		Method
	Small $n < 50$	**Medium** $50 \leq n < 100$	**Large** $n \geq 100$	**Medicine**	**Geology**	**Biologic**	**Other**	**Monte Carlo**	**Other**	$p < n$	$p > n$	
(Clarotto et al., 2021)	√				√			√		√		(α-IT)
(Bars-Cortina, 2022)			√	√				√			√	CLR
(Sisk-Hackworth et al., 2020)	√				√				√		√	CLR
(Smithson & Broomell, 2022)	√						√	√		√		ILR
(Templ et al., 2020)		√		√					√	√		ANNs
(Greenacre, 2022)			√	√				√			√	Box-Cox

2.2.2 Application of Compositional Data

This work shows that the linear models enjoys asymptotic false discovery rate control and can be extended to mixed-effect models for correlated microbiome compositional data and demonstrate the effectiveness (Zhou et al., 2021). The additive log-ratios of high-dimensional compositional data can provide a valid choice as transformed variables that are subcompositionally coherent, explaining 100% of the total log-ratio variance and coming measurably very close to being isometric (Greenacre et al., 2021). A reproducible vignette is provided for the application of selbal, a forward selection approach for the identification of compositional balances, and clr-LASSO and coda-LASSO, two penalized regression models for compositional data analysis, to enable researchers to fully leverage their potential in microbiome studies (Susin et al., 2020).

Analysis of clr transformation to reanalyze the next-generation sequencing and metabolomics data from a study investigating the effects of building material type, moisture and time on microbial and metabolomic diversity revealed novel relationships and stronger associations between sample conditions and microbial and metabolic community profiles (Sisk-Hackworth & Kelley, 2020). The relationship between time-of-day composition of exercise behavior and risk of death was explored using compositional data analysis, and hazard ratios for mortality were estimated based on Cox regression models (van Rosen et al., 2019).

This mini-review introduces microbiology researchers to compositional data analysis, data transformations, and various network theory methods, including static, temporal, sample-specific, and differential networks, aiming to reveal insights into biological phenomena and complex systems (Espinoza et al., 2020). A new

compositional data loss function (CD-trace) is proposed based on the D-trace loss. The sparse matrix estimator for direct interaction networks is defined as the minimization of the lasso penalized CD-trace loss under positive finite constraints (Yuan et al., 2019).

Table 2.2 shows the literature about the application of compositional data. From Table 2.2, the gaps are as follows:

(1) Compositional data is currently the subject of research focused on log-ratio transformations and predictions in various fields, with few articles discussing missing values.

(2) Not any paper published using the composition of employment industry in Taiyuan, China.

(3) Not any paper published using the composition of hospitalization expenses in Taiyuan, China.

(4) Not many application published using LASSO to deal with compositional data.

Table 2.2 Literature about the application of compositional data

Author(s) Year(s)	Sample size			Field of data set				Simulations		Data Dimension		Method
	Small $n < 50$	**Medium** $50 \leq n < 100$	**Large** $n \geq 100$	**Medicine**	**Geology**	**Biologic**	**Other**	**Monte Carlo**	**Other**	$p < n$	$p > n$	
(Zhou et al., 2021)			√	√				√		√		Linear models
(Greenacre et al., 2021)			√			√		√			√	PCA
(Susin et al., 2020)			√			√			√	√		CLR-LASSO
(Sisk-Hackworth & Kelley, 2020)			√			√			√		√	CLR
(van Rosen et al., 2019)		√					√	√		√		Cox
(Espinoza et al., 2020)			√	√		√		√		√		Linear models
(Yuan et al., 2019)		√				√	√		√	√		Loss function

2.3 High-dimensional data

High-dimensional data frequently appears in fields such as bioinformatics, biomedical research, econometrics, and machine learning. The term "high-dimensional" refers to a situation where the number of unknown parameters to be estimated is several times or more than the sample size. In traditional statistical studies, the number of observations (n) is usually greater than the number of variables (p) (Schintler, 2021). However, in high-dimensional data, the opposite is true, i.e., $n < p$.

Experimental results on the implementation of a quantum kernel classifier on real high-dimensional data from the field of cosmology using Google's general-purpose quantum processor Sycamore. By constructing a circuit equation that preserves the kernel magnitude that typically vanishes due to the exponential growth of the Hilbert space, and implementing error mitigation measures specifically tailored to the task of computing quantum kernels on recent hardware (Peters et al., 2021). The Elastic-net model is selected as the basic regularization model for processing high-dimensional sparse data, and a penalty factor is added to enhance its ability to retain key features and the rationality of penalty factor addition is confirmed. (Ma et al., 2023).

An R package is presented that provides a comprehensive set of tools for applying residual randomization in stochastic permutation procedures (RRPP) to linear models and provides comprehensive results for downstream analysis and mapping after model fitting using the lm.rrpp function (Collyer et al., 2018). A downscaling visualization algorithm is introduced-PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding), which consistently preserves

various patterns in the data, including continuous progression, branching, and clustering, and is applicable to a wide range of data types compared to other tools (Moon et al., 2019).

The protocol for analyzing high-dimensional cytometry data using FlowSOM, a clustering and visualization algorithm based on a self-organizing map, is described, which provides clearly annotated R code and an example dataset for inexperienced users (Quintelier et al., 2021). A feature screening procedure based on the correlation of distances between the marginal distributions of the covariates and the marginal distributions of the responses to the missing data is proposed, and the results show that the proposed method and algorithm perform well for linear models even when both the responses and the covariates are missing at random, and also work well for nonlinear models (Xia et al., 2023). A variable size co-evolutionary particle swarm optimization algorithm (VS-CCPSO) for feature selection is investigated, and the experimental results show that the algorithm has the ability to obtain a good subset of features, which suggests its competitiveness in dealing with high-dimensional feature selection problems (Song et al., 2020). This review examines in detail various feature extraction and feature selection methods are studied in detail and systematically compares several dimensionality reduction techniques for analyzing high dimensional data and overcoming data loss problems (Ray et al., 2021).

A new classifier ensemble method based on subspace enhancement (CESE) is proposed for high-dimensional data classification and comparative results indicate that the approach CESE outperforms different mainstream integrated systems (Xu et al., 2023). An Ensemble Classifier with Hybrid Feature Transformation for handling extremely complex dimensional data is proposed and proven to be efficient for selecting and extracting features from a very large dataset in a classification task

(Gunasundari & Arun, 2022) . This work develops a step-by-step scheme for full-spectrum high-dimensional data preparation for use in high-dimensional data analysis, with a focus on visualizing the impact of each step of data preparation using dimensionality reduction algorithms (Ferrer-Font et al., 2023).

By integrating Lasso regression and Ridge regression, an improved regularization term is proposed, which can be applied to datasets with different sparsities and can continuously improve the performance of existing benchmarks (Chai et al., 2022). The research aims to compare some methods of choosing the explanatory variables and the estimation of the parameters of the regression model, which are Bayesian Ridge Regression (unbiased) and the adaptive Lasso regressionmodel, using simulation (Muttaleb, 2023). By analyzing various factors and cholesterol levels, machine learning models can accurately predict an individual's likelihood of experiencing a heart attack, allowing healthcare providers to implement early intervention and preventive measures to reduce the risk of cardiovascular disease (Patil & Bhosale, 2023).

A fuzzy clustering feature selection method based on particle swarm optimization (PSOFS-FC) is proposed, and the experimental results show that the method can achieve excellent classification performance with relatively low classification accuracy. This indicates its superiority in handling high-dimensional unbalanced missing value data (Zhang et al., 2022). The accuracy of predictive models using features selected by the filtering method is analyzed, concluding that the simple variance filter outperforms all other filtering methods considered and identifying the correlation-adjusted regression score filter as a finer-grained alternative to fitting models with similar predictive accuracy (Bommert et al., 2021). The proposed combination of resampling-based least absolute shrinkage and