©Copyright 2023
Fang Nan

A Power Transformation-based Compositional Data Analysis Approach with Application to Physical Activity Epidemiology

Fang Nan

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2023

Committee:
Chongzhi Di, Chair
Jing Ma

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

A Power Transformation-based Compositional Data Analysis Approach with Application to Physical Activity Epidemiology

Fang Nan

Chair of the Supervisory Committee:
Chongzhi Di
Department of Biostatistics

Compositional data arise in many scientific fields, where relative proportions of different parts of a whole are basic units of data. An example is physical activity (PA) epidemiology, where one is often interested in composition of various PA intensity categories within 24-hour activity cycles based on objective measurements such as accelerometry. Although a few compositional data analysis approaches have been applied to modeling PA data, they have drawbacks that are often overlooked. In this master's thesis, we propose a power transformation-based framework for analyzing PA compositional data, which is more flexible and directly addresses the drawbacks of existing approaches. We first review current compositional data analysis approaches and their applications to PA data. Next, we present the proposed model for compositional data in the absence of zero values and investigate its theoretical properties, estimation, and inference. Moreover, we extend our power transformation-based model to account for compositional data in the presence of exact zeros. Two estimation strategies, constrained maximum likelihood estimation and modified likelihood procedures, are proposed. Extensive simulation studies were conducted to evaluate the finite sample properties of the proposed approaches. Finally, we applied these methods to study the compositional effects of sedentary behavior, light intensity PA and moderate to vigorous PA in relationship to health outcomes from the National Health and Nutrition Examination Survey (NHANES) data.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The past two years at UW have been an incredible journey for me, filled with countless experiences that have shaped me into who I am today. Throughout this time, I have eagerly embraced every opportunity that came my way, savoring the essence of each endeavor, both the challenges and the joys they brought. Regardless of the outcomes, I firmly believe that these experiences have bestowed upon me invaluable lessons that will resonate throughout my lifetime.

First and foremost, I would like to express my deepest appreciation to my advisor, Professor Chongzhi Di. Your guidance has been instrumental in shaping my ability to engage in rigorous academic research and think critically as a biostatistician. I am truly grateful for your insightful mentorship, which has broadened my perspective and equipped me with invaluable skills. Moreover, I am indebted to you for your unwavering encouragement and support during times of adversity and difficulty.

My heartfelt gratitude extends to the second reader on my committee, Professor Jing Ma. I am sincerely thankful for your invaluable comments and feedback on my thesis, which have undoubtedly contributed to its refinement. Additionally, I am immensely grateful for the opportunity to be involved in your captivating project on the dimensionality reduction of zero-inflated microbiome data. This experience has been nothing short of rewarding, and I am truly honored to have collaborated with you.

I would also like to extend my thanks to Professor Ken Rice for granting me the privilege to work alongside you on the Bayesian shrinkage estimates project. Your support and guidance have been invaluable throughout my studies in the department, and I am truly grateful for the enriching experiences I have gained through our collaboration.

Furthermore, I would like to express my heartfelt gratitude to my cherished friends,

whose unwavering support and camaraderie have been a constant source of strength and inspiration. Their presence in my life has made the challenges more bearable and the triumphs more meaningful.

Lastly, but most importantly, I want to express my deepest gratitude to my parents. Your constant presence, unwavering support, and unconditional love have been the cornerstone of my achievements. Without you, I would not have been able to accomplish all that I have.

Chapter 1

INTRODUCTION

In this chapter, we first present an introductory overview of the field of physical activity (PA) epidemiology and highlighted the recent advancements in this area over the past decades. Our discussion primarily focuses on the recent paradigm shift towards 24-hour activity cycles, from investigating each activity intensity/behavior separately to recognizing their interdependence. We also introduce the concept of compositional data analysis within the realm of physical activity epidemiology, encompassing fundamental principles that underpin this statistical approach. Moreover, we provide a brief introduction to the National Health and Nutrition Examination Survey (NHANES), which serves as the primary motivating study for methods development in this thesis.

1.1 Physical Activity Epidemology

Physical activity epidemiology focuses on investigating the relationship between levels and patterns of PA and their impact on health outcomes in populations (Wilmot et al. [2012], Caspersen [1989]). This field explores various aspects of PA, including different intensity of physical activity behaviors, such as sedentary behavior (SB), light intensity physical activity (LPA), and moderate to vigorous intensity physical activity (MVPA). Notably, accumulating evidence suggests that regularly engaging in MVPA per day is associated with improved health outcomes (Officer. [2013], Haskell et al. [2007]). Moreover, research has indicated that the time spent in LPA may also play a crucial role in preventing obesity (Kotz and Levine [2005]), whereas sedentary time has been consistently linked to negative health effects (Owen et al. [2009], Cliff et al. [2016], Tremblay et al. [2011]).

Although the investigation of PA behaviors is of great scientific interest, measuring PA behavior presents significant challenges due to its complexity (Caspersen et al. [1985], Skender et al. [2016]). Traditionally, one of the most common approaches is the use of PA questionnaires. These

questionnaires are capable of assessing all types of PA and can be employed in large sample sizes. However, due to the intricate and subjective nature of the information collected, PA questionnaires may be subject to substantial error and systematic bias (Ferrari et al. [2007], Neilson et al. [2008]).

1.2 Accelerometers

To address limitations inherent in PA questionnaires, motion sensors such as accelerometers have gained increasing popularity as a means of measuring physical activity in real-world settings (Yang and Hsu [2010]). Accelerometers are compact electronic devices that capture acceleration patterns associated with body movement, providing an objective assessment of locomotion duration and intensity (Westerterp [1999]). Currently, a wide range of accelerometers from various manufacturers is available in the market (Bai et al. [2016b]). The output of these devices is a high-resolution three-dimensional time series of accelerations expressed in gravitational units in the device's frame of reference. These raw data produced by accelerometers are transformed using various algorithms into PA summary metrics (Pedišić and Bauman [2015]). For example, the Actilife software converts raw accelerometry data collected by Actigraph accelerometers into counts per epoch (e.g., 1-minute), which provides a summary metric that indicates the aggregated movement intensity in each epoch (John and Freedson [2012]).

Accelerometers are now widely adopted by many large-scale epidemiological studies to collect information on participants' habitural PA patterns over a pre-specified time period (typically a few days to 1-2 weeks). A few examples include the National Health and Nutrition Examination Survey (NHANES) conducted from 2003 to 2006 and 2011-2014 (Varma et al. [2017], Urbanek et al. [2018]) and the Women's Health Initiative Objective Physical Activity and Cardiovascular Health Study (OPACH, LaCroix et al. [2017]).

1.3 Compositional Data Analysis

In the past, researchers often viewed various PA behaviors as separate entities, disregarding their interdependent nature (Bansil et al. [2011], Haskell et al. [2007]). However, it is crucial to recognize that the combined duration of physical activity behaviors such as SB, LPA, and MVPA contributes to the overall waking hours (Maher et al. [2014]). Moreover, incorporating sleep as a

distinct category alongside these physical activity behaviors guarantees a comprehensive depiction of the complete 24-hour activity cycle. Consequently, alterations in one behavior will invariably influence the remaining behaviors (Chastin et al. [2015]). In addition, researchers become increasingly aware of the limitations associated with considering physical activity behaviors as individual and independent components (Pedišić [2014], Pearson [1897]). This recognition has been particularly driven by the advancements in accelerometers, which enable more detailed and precise measurements of various PA behaviors. As a result, there is a shift towards a new paradigm that considers these behaviors as mutually exclusive and exhaustive parts of a 24-hour day, i.e., the 24-hour Activity Cycle (24-HAC) (Rosenberger et al. [2019]).

To tackle the aforementioned challenges in physical activity epidemiology, the implementation of compositional data analysis emerges as an inherent statistical approach (Janssen et al. [2020]). Compositional data is a special type of multivariate data where the values of each observational vector are non-negative and sum to a constant, usually equal to 1 for convenience purposes. They have been commonly encountered in various scientific fields including microbiome studies (Tsilimigras and Fodor [2016]), nutrition (Ros-Freixedes and Estany [2014]), geochemistry (Montero-Serrano et al. [2010], Buccianti et al. [2014]), politics (Honaker et al. [2002]), and behavioral biology (Pierotti et al. [2009]). In the physical activity context, if we denote the average daily time (in minutes) spent in each PA category for subject i as $(t_{i,1},...,t_{i,D})$, where $t_{i,D}$ is the person's average sleep time and $t_{i,d}, k = 1,...,D-1$ are the time spent in the PA categories while awake, naturally $\sum_d t_{i,d} = 1440$. The vector $\mathbf{t}_i = (t_{i,1},...,t_{i,D})^T$ belongs to the compositional subspace of $\mathbb{R}^D_+$, the non-negative D-dimensional real Euclidean space. Without loss of generality, consider $x_{i,d} = t_{i,d}/1440$ as the daily average time proportion spent in PA category d, the support of $\mathbf{x}_i = (x_{i,1},...,x_{i,D})^T$, termed simplex, is given by:

$$\mathbb{S}^{D-1} = \left\{ (x_{i,1},...,x_{i,D})^t | x_{i,d} \geq 0, \sum_k x_{i,d} = 1 \right\}. \tag{1.1}$$

The data belonging to the simplex is termed compositional data. Thus due to the smu to 1 constraint, the traditional linear regression with N observations that uses all components from a composition as covariates will have a singular design matrix $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_N)^T$.

Various techniques have been developed for analyzing compositional data, both with and with-

out consideration of the simplex structure (Alenazi [2021]). Traditional methods designed for Euclidean data cannot be directly adapted to compositional data due to the constrained nature of the simplex. In the field of physical activity epidemiology, an isotemporal substitution paradigm has been developed as a novel analytic model to examine the time-substitution effects of one activity for another (Mekary et al. [2013]). Additionally, log-ratio transformations, as proposed by Aitchison [2003], offer a means of mapping the data to the Euclidean space. This transformation eliminates the unit sum constraint and enables the application of standard multivariate techniques. An example of such an application can be seen in the work by Chastin et al. [2015], who was the first to employ this model in physical activity research. The study investigated the combined effect of different behaviors on obesity and cardio-metabolic health markers. However, these transformations, along with some other suggested parametric models, prove inadequate when observations lie on the boundaries of the simplex (Alenazi [2021]). We will provide a review of existing methods and discuss their pros and cons in Chapter 2.

1.4 The National Health and Nutrition Examination Survey (NHANES)

The National Health and Nutrition Examination Survey (NHANES) is a pivotal program designed to evaluate the health and nutritional well-being of adults and children in the United States. What sets this study apart is its employment of both interviews and physical examinations to collect data. NHANES employs a sophisticated sampling design to generate a representative sample of the US civilian non-institutionalized population. This includes an in-person home interview, as well as a visit to a mobile examination center, where laboratory data are collected. The study protocols have received approval from the ethics review board of the Centers for Disease Control and Prevention in Atlanta, Georgia, and all subjects have provided informed consent.

For this thesis, we analyzed data from the 2005-2006 study cycle of NHANES (National Center for Health Statistics [2005]). The sample comprised 4,979 respondents aged 20 years or older, who were interviewed and examined with response rates of 74.4% and 71.5%, respectively. Exclusions were made for respondents with diagnosed sleep disorders, current pregnancy, lactation, or insulin use (N = 1,945), insufficient valid accelerometry data, missing self-reported sleep duration, covariate, or biomarker data. This resulted in a sample of 2,185 adults for the full analysis

(52.9% of eligible subjects) and 923 adults for the fasting subsample (22.3% of eligible subjects), which were representative of the total population. The physical activity and sedentary behavior of the study participants were measured objectively using the ActiGraph accelerometer (model 7164; ActiGraph LLC, Pensacola, FL). Participants were instructed to wear the device on their right hip for 7 consecutive days during all waking hours, except for bathing or swimming, and an elastic belt was used to secure the accelerometer to their body. The device was pre-programmed to record data in 1-minute intervals, ensuring a high level of accuracy in the measurement of activity levels and sedentary behavior.

Numerous studies have utilized the NHANES 2005-2006 data to investigate health-related behaviors. For instance, Buman et al. [2014] investigate the association between reallocating time to sleep, sedentary behavior, or active behaviors with biomarkers, and they find that MVPA may be the most potent time-dependent behavior that enhances health, with additional benefits from LPA and sleep duration when reallocated from sedentary time. Vallance et al. [2011] identify a lower likelihood of depression associated with increasing MVPA and decreasing sedentary time, especially among overweight or obese adults. Similarly, Choi and Ainsworth [2016] find that individuals with the highest step count exhibited a healthier eating profile and better serum vitamin levels compared to their less active counterparts. Conversely, those with the lowest step count had an increased likelihood of having metabolic syndrome and its risk components. These studies underscore the importance of engaging in regular physical activity and avoiding prolonged sedentary behavior for improved health outcomes, as supported by statistical analysis of the NHANES 2005-2006 data.

1.5 Outline of Thesis

This thesis is organized into seven chapters. Chapter 2 provides a review of existing methodologies for compositional data analysis, laying the foundation for the subsequent chapters. In Chapter 3, we introduced power transformation-based regression models for compositional data in the absence of exact zero values. Chapter 4 extends the model to account for zero compositions, where estimation procedures, constrained and modified maximum likelihood approaches, were proposed. Simulation results are presented in Chapter 5, providing empirical evidence to demonstrate finite

sample properties of the proposed approaches. Chapter 6 demonstrates the real-world applications of the proposed methodology to NHANES data. Chapter 7 includes concluding remarks and directions for future research.

Chapter 2

REVIEW OF COMPOSITIONAL DATA ANALYSIS METHODOLOGY

In this section, we begin by introducing isotemporal substitution regression, a popular technique for modeling compositional PA data. Subsequently, we conduct an overview of log-ratio based approaches for analyzing compositional data, with a particular emphasis on the isometric log-ratio transformation. Next, we introduce a power transformation family, which served as a more flexible extension of the aforementioned methodologies.

2.1 Isotemporal Substitution Model

A modified approach to performing linear regression using compositional data as covariates is referred as isotemporal substitution model (ISM, Mekary et al. [2009]), which only uses $D-1$ components from $\mathbf{x}_i$ to ensure the design matrix has full column rank. If we delete the d-th column of the full design matrix $\mathbf{X}$, denoted as $\mathbf{X}^{(-d)}$, the regression could be written as:

$$\mathbf{Y} = \beta_0^{(-d)} + \mathbf{X}^{(-d)}\beta^{(-d)} + \varepsilon,$$

where ε is the error term. The interpretation of the j-th regression coefficient $\beta_j^{(-d)}$ is the expected change in Y while substituting 1 unit of x_d by x_j. Thus by repeating fit isotemporal regression with different $\mathbf{X}^{(-d)}$ as design matrix, one could estimate the substitution effect between any pairs (x_j, x_d) from the composition $\mathbf{x} = (x_1, ..., x_D)$ (Dumuid et al. [2019]).

However, there are several disadvantages of the isotemporal regression approach (Weipeng et al. [2023]). The linear model formulation assumes that the substitution effects between compositional pairs remain consistent across the simplex. For example, when applied to PA analysis, it automatically indicates that on average spending 30 more minutes in MVPA from subtracting 30 minutes in SB daily will have the same effect on the health outcome for both very active and very

sedentary populations, which is unlikely always true.

2.2 Log-ratio Based Regression

Over the past 40 years, significant advancements have been made in the field of compositional data analysis, particularly through the development of log-ratio based approaches. These approaches have proven invaluable in handling compositional data, with the ability to effectively address the challenges posed by the simplex constraints. The additive log-ratio (ALR) transformation, introduced by Aitchison [1982] and Atchison and Shen [1980], allows for the escape of unit sum constraint by mapping data onto the Euclidean space, $\mathbb{R}^{D-1}$, where standard multivariate techniques can be applied. However, one possible drawback is that it treats components asymmetrically, leading to varying interpretations of the analysis depending on the chosen common divisor. In response, Aitchison [1983] proposed the centered log-ratio (CLR) transformation, which treats the data symmetrically while still allowing it to lie within the Euclidean space, but introduces the zero-sum constraint. The singularity problem resulting from the CLR-transformation can be resolved by multiplying CLR with the Helmert sub-matrix $\mathbf{H}$. The Helmertized CLR transformed data can then be mapped onto a D–1-dimensional real space. This transformation is known as the isometric log-ratio (ILR) transformation, as termed by Barceló-Vidal et al. [2001] and Egozcue et al. [2003]. The term "isometric" refers to the fact that the distances between two compositional vectors remain the same before and after the multiplication by the Helmert sub-matrix. It is worth noting that any orthonormal matrix can be used as a substitute for the Helmert sub-matrix, as demonstrated by Egozcue and Pawlowsky-Glahn [2005].

ILR-transformation has been implemented for regression analysis with compositional PA data. This transformation is based on orthogonalizing the basis of the Euclidean space composed of CLR-transformed compositional data $\mathbf{c}_i = \text{CLR}(\mathbf{x}_i) = (c_{i,1}, ..., c_{i,D})^T$, with d-th element:

$$c_{i,d} = \log\left(\frac{x_{i,d}}{\prod_d x_{i,d}^{1/D}}\right) \text{ for } d = 1, ..., D.$$

The ILR transformation is defined as $\mathbf{z}_i = \text{ILR}(\mathbf{x}_i) = \mathbf{c}_i \cdot \Psi^t$, where Ψ is a $D \times D$ orthonormal matrix with its first row deleted, for example, the Helmert matrix (Egozcue et al. [2003]). With the

ILR transformed compositional data $\mathbf{Z}_{N\times(D-1)} = (\mathbf{z}_1, ..., \mathbf{z}_N)^T$ as the design matrix, which has full column rank, the linear regression:

$$\mathbf{Y} = \beta_0^{ILR} + \mathbf{Z}\beta^{ILR} + \varepsilon$$

is denoted as ILR-transformed regression for regressing $\mathbf{Y}$ on compositional data $\mathbf{X}$, where ε is the error term. The direct interpretation of regression coefficients β^{ILR} is however not trivial. For PA analysis, it is easier to graphically illustrate the substitution effect between PA category pairs using the difference in predicted outcome $\mathbf{Y}$ as shown in Chastin et al. [2015].

One major disadvantage of the ILR-transformation based approach is rooted in its dependence on log-ratio quantities. When one or more observed components are zero in the compositional data, both CLR and ILR-transformation could not be performed. The zero observation in PA measurements is not rare, especially in the highest activity category from the less active population. There could be potential bias from either discarding these subjects with zero PA components or assigning a small number in place of the zeros. Discarding subjects with zeros not only sacrifices statistical efficiency but also disregards the contribution of the specific activity pattern group to the model, while replacing zeros with a small value may produce significant bias in the regression coefficient estimation since the log-ratio transformation has near-infinity leverage when the zero-component is in the denominator during transformation. Thus either approach distorts the ILR transformation regression model fitting heavily.

2.3 Power Transformation-based Regression

Tsagris et al. [2011] introduces the α-transformation, which incorporates a power parameter α. In the context of this thesis, we have chosen to refer to this transformation as the power transformation to prevent any potential confusion. This power transformation has found successful applications in discriminant settings, as evidenced by the studies conducted by Ankam and Bouguila [2018] and Tsagris et al. [2016].

As an extension of the ILR-transformation for compositional data, the power transformation is in a spirit similar to the Box-Cox transformation for closer to the Gaussian distribution of trans-

formed data. Instead of performing the CLR-transformation, let

$$\mathbf{u}_i^{(\alpha)} = \left(\frac{x_{i,1}^{\alpha}}{\sum_d x_{i,d}^{\alpha}}, ..., \frac{x_{i,D}^{\alpha}}{\sum_d x_{i,d}^{\alpha}} \right)^T, \tag{2.1}$$

where α is a real-valued parameter within $[0, 1]$. The power transformation can be defined as:

$$\mathbf{z}_i^{(\alpha)} = \left(\frac{D\mathbf{u}_i^{(\alpha)}}{\alpha} - \frac{\mathbf{1}}{\alpha} \right) \Psi^T, \tag{2.2}$$

where $\mathbf{1}$ is a D-dimensional vector with all elements equal to 1, and Ψ could be the same Helmert matrix with its first row deleted as used in ILR-transformation. One interesting property of $\mathbf{z}_i^{(\alpha)}$ is that as $\alpha \to 0$ the power transformed data $\mathbf{z}_i^{(\alpha)}$ converges to ILR-transformed data $\mathbf{z}_i$, while as $\alpha = 1$ the $\mathbf{z}_i^{(\alpha)}$ is just an orthogonalized version of original compositional data $\mathbf{x}_i$ with parallel translation. Consider power transformed compositional data $\mathbf{Z}_{N\times(D-1)}^{(\alpha)} = (\mathbf{z}_1^{(\alpha)}, ..., \mathbf{z}_n^{(\alpha)})^T$ as the design matrix, the regression:

$$\mathbf{Y} = \beta_0 + \mathbf{Z}^{(\alpha)}\beta + \varepsilon,$$

is denoted as power transformation-based regression (PTR), where ε is the error term. Naturally, the ILR-transformed regression and isotemporal regression are just specific cases of the PTR when α is 0 or 1. For α being an interior point within [0,1], PTR is related to but different from these two specific cases and is a blend of both regression models while avoiding the drawbacks of each. One significant advantage of PTR over the ILR-transformed regression is that zero-count components no longer need to be treated as long as $\alpha \neq 0$. Note, α needs to be estimated but it does not require as much computational resource as the fully non-parametric regression models. The model coefficient interpretation, similar to the ILR-transformed regression model, is non-trivial but could be illustrated using predicted model outcomes versus substitution quantities between PA category pairs.

As previously mentioned, the behavior of PTR varies depending on whether the compositional data contains zeros or not. In the case where compositional data does not contain any zeros, the parameter α can assume any value within the range of [0, 1]. This allows for the utilization of standard estimation procedures during implementation. Conversely, when the data includes zero compositions, α cannot be set to 0. In such instances, the power transformation converges to the

ILR-transformation as α approaches 0, necessitating the adoption of distinct estimation procedures. Consequently, in the following chapters, we will address these two scenarios separately. First, we will discuss the estimation procedures for modeling non-zero compositional data, and then we will delve into the more intricate situation of modeling compositional data containing zeros.

Chapter 3

POWER TRANSFORMATION-BASED REGRESSION FOR COMPOSITIONAL DATA WITHOUT ZERO VALUES

In this section, we begin by exploring the properties of the power transformation. Next, we focus on utilizing the PTR to model compositional physical activity data without zero values and consider estimation procedures based on Maximum Likelihood Estimation and Generalized Cross-Validation. Lastly, we propose a bootstrap-based confidence interval to predict outcomes while accounting for uncertainty due to the estimation of the tuning parameter α.

3.1 Orthogonality for the power transformation

The orthogonality of the ILR-transformation and the construction of ILR-associated orthonormal bases were investigated by Egozcue et al. [2003]. However, the existing literature has not explored the orthogonality of the power transformation. Given the relationship between the power transformation and the ILR-transformation, as well as the previous work conducted by Egozcue et al. [2003], we aim to examine the orthogonality of the power transformation in this section. To achieve this, we begin by introducing an inner product based on the power transformation using the definitions provided in (2.1) and (2.2).

Definition 1. *For any two compositions* $\mathbf{x}$ *and* $\mathbf{y}$ *in* S^D*, the inner product based on power transformation is*

$$\langle \mathbf{x}, \mathbf{y} \rangle_\alpha = \frac{1}{\alpha^2} \sum_{i=1}^{D} \left(\frac{Dx_i^\alpha}{\sum_{j=1}^{D} x_j^\alpha} - 1 \right) \left(\frac{Dy_i^\alpha}{\sum_{j=1}^{D} y_j^\alpha} - 1 \right), \tag{3.1}$$

which induces a norm in S^D *in the standard way*

$$||\mathbf{x}||_\alpha^2 = \langle \mathbf{x}, \mathbf{x} \rangle_\alpha .$$

It is important to note that as α approaches 0, the power transformation converges to ILR transformation. This is demonstrated by the fact that the inner product defined in (3.1) converges to the inner product defined based on ILR-transformation:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{ILR} = \sum_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \ln \frac{y_i}{g(\mathbf{y})},$$

where $g(\mathbf{x}) = [x_1 x_2 \dots x_D]^{1/D}$ is the geometric mean of the components of $\mathbf{x}$. This highlights the connection between the two transformations and suggests that the power transformation can be seen as a generalization of the ILR transformation, allowing for more flexibility in compositional data analysis.

In order to obtain an orthonormal basis for the power transformation, we selected a set of $D-1$ linearly independent vectors. Let $\mathbf{v}_1$, $\mathbf{v}_2$,..., $\mathbf{v}_{D-1}$ defined as $\mathbf{v}_i = [0,\dots,0,1,-1,0,\dots,0]$, with the first non-null element being placed in the i-th column. Therefore, the Gram–Schmidt procedure, with respect to the ordinary Euclidean inner product, can be applied to $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{D-1}$ and we can obtain the following result.

Proposition 1. *Let $\boldsymbol{u}_i \in \mathbb{R}^D$, $i = 1,\dots,D-1$ be the vectors:*

$$\boldsymbol{u}_i = \sqrt{\frac{i}{i+1}} \left[\frac{1}{i}, \dots, -1, 0, \dots, 0 \right], \tag{3.2}$$

where first i elements share the same value, $\sqrt{\frac{i}{i+1}}\frac{1}{i}$ and the $(i+1)$-th element is $-\sqrt{\frac{i}{i+1}}$ and remaining elements are zero. The vector $\boldsymbol{u}_i$'s are orthonormal with respect to the ordinary Euclidean inner product defined in $\mathbb{R}^D$ and constitute a basis of (D-1)-dimensional linear subspace.

With the previous results, we can obtain an orthonormal basis for the power transformation.

Theorem 1. *Let $\boldsymbol{e}_i$, $i = 1,\dots,D-1$, be the following compositions in S^{D-1}:*

$$\boldsymbol{e}_i = \left(\left[\alpha \sqrt{\frac{1}{i(i+1)}} + 1 \right]^{1/\alpha}, \dots, \left[\alpha \sqrt{\frac{1}{i(i+1)}} + 1 \right]^{1/\alpha}, \left[-\alpha \sqrt{\frac{i}{i+1}} + 1 \right]^{1/\alpha}, 0, \dots, 0 \right), \tag{3.3}$$

where the first i elements are the same. The vector $\boldsymbol{e}_i$ are orthonormal with respect to the inner product defined in (3.1) and they are a basis of S^{D-1}.

Proof. The inner product between $\mathbf{e}_i$ and $\mathbf{e}_j$ can be computed using (3.1)

$$\left\langle \mathbf{e}_i, \mathbf{e}_j \right\rangle_\alpha = \frac{1}{\alpha^2} \sum_{k=1}^{D} \left(\frac{D\mathbf{e}_{ik}^\alpha}{\sum_{m=1}^{D} \mathbf{e}_{im}^\alpha} - 1 \right) \left(\frac{D\mathbf{e}_{jk}^\alpha}{\sum_{m=1}^{D} \mathbf{e}_{jm}^\alpha} - 1 \right) = \frac{1}{\alpha^2} \sum_{k=1}^{D} \alpha^2 u_{ik} u_{jk} = \mathbf{u}_i \mathbf{u}_j' = 0$$

for $i \neq j$ due to the orthogonality of $\mathbf{u}_i$, $\mathbf{u}_j$ in $\mathbb{R}^D$ and the fact that $\sum_{m=1}^{D} \mathbf{e}_{jm}^\alpha = D$. Normalization to the unity of $\mathbf{e}_i$ follows from the same expression by taking $i = j$. □

A transformation between S^{D-1} and $\mathbb{R}^{D-1}$ can be obtained in a standard way by using the power transformation associated orthonormal basis, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$. We seek the power transformation such that $\alpha(\mathbf{e}_i) = \vec{e}_i$ for $i = 1, 2, \ldots, D-1$, where $\vec{e}_i$ is the i-th vector of the canonical basis in $\mathbb{R}^{D-1}$. This desired transformation is defined as follows:

Definition 2. *For any composition $\boldsymbol{x} \in S^{D-1}$, the power transformation associated with the orthonormal basis in S^{D-1}, $\boldsymbol{e}_i$, $i = 1, 2, \ldots, D-1$, is the transformation from S^D to $\mathbb{R}^{D-1}$ given by*

$$\mathbf{z}^{(\alpha)} = \alpha(\mathbf{x}) = [\langle \mathbf{x}, \mathbf{e}_1 \rangle_\alpha, \langle \mathbf{x}, \mathbf{e}_2 \rangle_\alpha, ..., \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_\alpha]. \tag{3.4}$$

3.2 Estimation Procedures

Consider the power transformed compositional data $\mathbf{Z}_{N\times(D-1)}^{(\alpha)} = (\mathbf{z}_1^{(\alpha)}, ..., \mathbf{z}_n^{(\alpha)})^T$ as the design matrix, defined in (2.2), and $\mathbf{Y}$ as the outcome, the regression:

$$\mathbf{Y} = \beta_0 + \mathbf{Z}^{(\alpha)}\beta + \varepsilon, \tag{3.5}$$

is denoted as PTR, where ε is the error term and α is a tuning parameter that controls the type of transformation and β is the vector of regression coefficients. Once α is estimated, estimates of β can be obtained straightforwardly by least squares. Thus, we focus on the estimation of α. Currently, several well-established methods exist for estimating α in the power transformation-based model. It is crucial to note, however, that the choice of criterion and approach for estimating α can vary depending on the specific application of the power transformation. For instance, if the goal is to obtain a spatial structure in the resulting vector $\mathbf{z}^{(\alpha)}$ that closely approximates a Gaussian-distributed random vector, one could choose α that minimizes the Kullback-Leibler divergence between $\mathbf{z}^{(\alpha)}$ and the desired distribution (Tsagris [2015]). However, when applying a regression

model to PA data, which is the scenario that we focus on here, one might choose α based on considerations related to predictive performances.

In this section, we propose two estimators for α: one is derived from the likelihood and the other is based on cross-validation (CV). By assuming the regression residuals ε follow a normal distribution with mean 0 and variance σ^2, the full data likelihood can be expressed as:

$$\mathcal{L}(\alpha,\beta,\sigma^2) = \sum_{i=1}^{N} log\left(\Phi\left[\frac{y_i - (1,\mathbf{z}_i^{(\alpha)T})^T\beta}{\sigma}\right]\right), \tag{3.6}$$

where N is the sample size, Φ represents the density function of a standardized Gaussian distribution. $\mathcal{L}(\alpha,\beta,\sigma^2)$ denotes the logarithm of the profile likelihood, where $\beta = \beta(\alpha)$ and $\sigma^2 = \sigma^2(\alpha)$, indicating $\ell(\alpha,\beta,\sigma^2)$ is actually a function of α. Once the likelihood is formulated, we can obtain the MLE of α

$$\hat{\alpha} = \underset{\alpha\in[0,1]}{\operatorname{argmax}}\,\mathcal{L}(\alpha,\beta,\sigma^2). \tag{3.7}$$

Then the estimator of β and σ can be obtained through the ordinary least squares techniques based on (3.6) and (3.7):

$$\hat{\beta}(\hat{\alpha}) = \left(\mathbf{Z}^{(\hat{\alpha})^T}\mathbf{Z}^{(\hat{\alpha})}\right)\mathbf{Z}^{(\hat{\alpha})T}\mathbf{Y}, \tag{3.8}$$

$$\hat{\sigma}^2(\hat{\alpha}) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{z}_i^{(\hat{\alpha})}\hat{\beta}(\hat{\alpha})\right)^2. \tag{3.9}$$

We also explored an alternative cross-validation (CV) based estimation approach, which aimed to minimize the Mean squared prediction error (MSPEs) between the regression fit and the observed values of $\mathbf{Y}$, over a pre-specified range of potential α values. The K-fold CV is usually employed for its balance between estimation accuracy and computational efficiency, by evenly separating the full data into K parts with an equal sample size. In each iteration, one part of the data $(\mathbf{Y}^k,\mathbf{X}^k), k = 1,\ldots,K$ is used as the testing data, while the other $K-1$ parts $(\mathbf{Y}^{(-k)},\mathbf{X}^{(-k)})$ are used as the training data, fitted with α taken from a set of values. A special case of K-fold CV when $K = N$ is referred to as leave-one-out CV, which is more robust but computationally intensive than a reasonably small K. Empirically 5 to 10 folds often yield negligible performance penalty compared to the leave-one-out CV (James et al. [2013]).

One could further accelerate the computation of MPSE via generalized cross-validation (GCV). The MSPE for a given α of GCV is defined as:

$$\text{MPSE}(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i(\alpha)}{1 - tr(H^\alpha)/D}\right)^2, \tag{3.10}$$

where $H^\alpha = (1, \mathbf{Z}^{(\alpha)})((1, \mathbf{Z}^{(\alpha)})^T(1, \mathbf{Z}^{(\alpha)}))^{-1}(1, \mathbf{Z}^{(\alpha)})^T$ is the hat matrix, $\hat{y}_i(\alpha) = (1, \mathbf{z}_i^{(\alpha)T})^T\beta(\alpha)$ is the fitted outcomes. Given that the performance of CV-type estimators, including K-fold CV, leave-one-out CV, and GCV, are relatively similar, we only focus on the performance of GCV-based estimators in the subsequent studies:

$$\hat{\alpha}_{GCV} = \underset{\alpha\in[0,1]}{\text{argmin}}\ \text{MPSE}(\alpha). \tag{3.11}$$

3.3 Asymptotic Properties

In this section, we study the asymptotic properties of the regression coefficients $\hat{\beta}$ and transformation parameter $\hat{\alpha}$. We first focused on the MLE. To derive the asymptotic distributions, we used the notation $\theta = (\beta, \alpha, \sigma^2)$, and express the log-likelihood for a single sample is

$$\ell(y, \mathbf{x}; \theta) = \log\left(\Phi\left[\frac{y - (1, \mathbf{z}^{(\alpha)T})^T\beta}{\sigma}\right]\right),$$

where $\mathbf{z}^{(\alpha)}$ is the power transformed $\mathbf{x}$ with a specific α. Thus, the log-likelihood of the complete data can be expressed as follows

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}; \theta) = \sum_{i=1}^{N}\ell(y_i, \mathbf{x}_i; \theta) = \sum_i log\left(\Phi\left[\frac{y_i - (1, \mathbf{z}_i^{(\alpha)T})^T\beta}{\sigma}\right]\right), \tag{3.12}$$

where $\mathbf{Y} = (y_1, y_2, \ldots, y_N)$ is the outcomes, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^T$ is the covariates. Then we can calculate the Fisher information matrix as

$$\mathbf{I}_N(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\mathcal{L}(y, \mathbf{x}; \theta)\right],$$

and define $I(\theta) = \lim_{N\to\infty} I_N(\theta)/N$.

We would like to discuss the asymptotic properties of the transformation parameter $\hat{\alpha}$ under two situations: (1) the true value of α, α_0, is an interior point within [0,1]; (2) α_0 is on the boundary

of its parameter space (e.g., $\alpha = 0,1$). When α_0 is an interior point of its parameter space, the asymptotics of the MLE and log-ratio test (LRT) follow standard large sample theory, as stated below.

Proposition 2. *When the true value of α, α_0, is an interior point of the interval [0,1], the asymptotic distribution of $\hat{\alpha}$ can be expressed as*

$$\sqrt{N}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22}) \tag{3.13}$$

where $\theta_0 = (\alpha_0, \beta_0, \sigma_0^2)$ with $\beta_0 = \beta_0(\alpha_0), \sigma_0^2 = \sigma_0^2(\alpha_0)$ as given in (3.8) and (3.9). $[\boldsymbol{I}(\theta_0)^{-1}]_{ij}$ is the entry at i-th row and j-th column of the matrix $\boldsymbol{I}(\theta_0)^{-1}$.

Proposition 3. *If α_0 is an interior point of the interval [0,1], the LRT statistic for testing the null hypothesis $H_0 : \alpha = \alpha_0$ versus alternative hypothesis $H_1 : \alpha \neq \alpha_0$ is given by*

$$\lambda_L = -2[\ell(\theta_0) - \ell(\hat{\theta})], \tag{3.14}$$

where $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ is the MLE of $\theta = (\alpha, \beta, \sigma^2)$. Under the null hypothesis, the test statistic λ_L converges to a Chi-squared distribution with 1 degree of freedom

$$\lambda_L \xrightarrow{d} \chi_1^2. \tag{3.15}$$

When α_0 is on the boundary of its parameter space, at 0 or 1, however, standard regularity conditions are violated and as a result, the asymptotics for the MLE and LRT do not apply anymore. Instead, we apply asymptotic theories for MLE under boundary conditions (Self and Liang [1987]).

Proposition 4. *When α_0 is on the boundary of its parameter space [0,1], the asymptotic distribution of $\hat{\alpha}$ has the following form:*

- *if $\alpha_0 = 0$, $\sqrt{N}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N^+(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$, where $N^+(0, [\boldsymbol{I}_n^{-1}]_{22})$ represents $N(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$ being left truncated at 0.*

- *if $\alpha_0 = 1$, $\sqrt{N}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N^-(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$, where $N^-(0, [\boldsymbol{I}_n^{-1}]_{22})$ represents $N(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$ being right truncated at 0.*

Proposition 5. *If $\alpha_0 = 1$, which is on the boundary of its parameter space, the LRT statistic for testing the null hypothesis $H_0 : \alpha = \alpha_0$ versus alternative hypothesis $H_1 : \alpha \neq \alpha_0$ is given by*

$$\lambda_L \xrightarrow{d} \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_0^2, \tag{3.16}$$

where $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_0^2$ is the 50:50 mixture of Chi-squared distributions with 1 and 0 degrees of freedom, respectively.

We can draw inferences regarding the MLE $\hat{\alpha}$ based on the results above. On the other hand, given the intricate and not yet fully comprehended nature of GCV-based estimators, it remains challenging to make direct inferences on such estimators. Nonetheless, recent literature has seen advancements in this area. For instance, Bates et al. [2021] explored inference for cross-validation, proposing an estimator for the mean squared error (MSE) of the cross-validation point estimate, as well as a nested CV scheme that exhibits consistently superior coverage in contrast to naive CV confidence intervals. Bayle et al. [2020] developed central limit theorems for cross-validation and consistent estimators of its asymptotic variance under weak stability conditions on the learning algorithm. These provide asymptotically-exact confidence intervals for K-fold test error and hypothesis tests to determine whether one learning algorithm has a smaller K-fold test error than another. However, it is important to highlight that these innovative methodologies might not be directly applicable to our specific scenario. The reason behind this lies in the fact that these methodologies predominantly focus on the K-fold CV technique rather than the GCV technique, which is an accelerated version of the K-fold CV. Consequently, further research in this domain remains necessary.

3.4 Bootstrap-based Confidence Interval for Predicted Outcomes

In the context of regression frameworks utilizing power transformation, the distribution of the design matrix $\mathbf{Z}^{(\alpha)}$ can often be difficult to determine due to its dependency on the value of α. As a result, the distribution of the regression residues may not be Gaussian. In order to overcome this challenge, we turn to the use of a bootstrap-based method to construct prediction confidence intervals, which do not require distributional assumptions for the residuals. The key advantage of

the bootstrap approach is that it is asymptotically invariant to the observation distribution, making it a powerful tool for constructing valid confidence intervals under various settings.

The confidence interval for the predicted outcome y_{new} from the input covariate $\mathbf{x}_{new}$ is computed following Stine [1985], which is based on the bootstrap procedure for constructing prediction interval from the linear regression model. We briefly describe the algorithm below:

1. Fit the PTR model by either GCV or maximum likelihood approach and get normalized residuals $(r_1, ..., r_N)$ from fitted data $\hat{y}_i$ with $r_i = (\hat{y}_i - y_i)/\sqrt{1-h_i}$, where $h_i = H^{\hat{\alpha}}_{ii}$ is the i-th diagonal element of the hat matrix H^{α}.

2. Get bootstrap samples $(r^*_1, ..., r^*_N)$ by sampling with replacements from $(r_1, ..., r_N)$.

3. Construct bootstrap samples of predicted outcome $(y^*_1, ...y^*_N)$ via $y^*_j = \hat{y}_j + r^*_j$, $j = 1, \ldots, N$.

4. Fit the regression model with $(y^*_j, \mathbf{x}_j)$, $j = 1, \ldots, N$, and denote the resulting model coefficient as $\tilde{\alpha}^*$ and $\tilde{\beta}^*$.

5. Obtain normalized residuals $(\tilde{r}^*_1, ..., \tilde{r}^*_N)$ from the new regression model obtained in step 4, similar to step 1.

6. With a new regression input x_{new}, obtain $\hat{y}_{new}$ and $\hat{y}^*_{new}$ from the original (step 1) and new regression model (step 4) respectively and generate prediction bootstrap sample by

$$r^*_{boot} = (\hat{y}_{new} - \tilde{y}^*_{new}) + \tilde{r}^*_k$$

   where k is randomly chosen from $(1, ..., N)$.

7. Repeat the above step 1-6 for N_{boot} times to obtain bootstrap samples $(r^*_{boot,1}, ..., r^*_{boot,N_{boot}})$ of the prediction error.

Let $r^*_{\alpha/2}$ and $r^*_{1-\alpha/2}$ be the lower and upper empirical $\alpha/2$ quantiles of $(r^*_{boot,1}, ..., r^*_{boot,N_{boot}})$, and the $1-\alpha$ level prediction confidence interval for y_{new} is constructed as:

$$(\hat{y}_{new} + r^*_{\alpha/2}, \hat{y}_{new} + r^*_{1-\alpha/2}).$$

By following the above algorithm, we can construct prediction confidence intervals that are robust to the distributional assumptions, making them applicable in a wide range of regression settings. Further assessments regarding the coverage probability of the bootstrap-based confidence intervals will be provided in the simulation sections.

The interpretation of model coefficients in PTR presents a non-trivial challenge due to the regression coefficients being associated with the transformed covariates rather than the PA behaviors themselves. Nonetheless, this complexity can be effectively elucidated by examining the predicted outcomes of the model compared to the substitution quantities observed between pairs of PA categories. To illustrate this concept, let us consider the substitution effects between the SB and LPA categories.

1. Fit the PTR model by either GCV or maximum likelihood approach.

2. Select a new observation, denoted as PA_0 = (SB_0, LPA_0, $MVPA_0$).

3. Utilize the fitted PTR model to calculate the predicted outcome with PA_0, denoted as P_0.

4. Keep the MVPA constant while manipulating the time allocation between SB and LPA. This adjustment can be achieved by setting a time change δ such that $SB_1 = SB_0 - \delta$ and $LPA_1 = LPA_0 + \delta$. Consequently, we obtain a new observation, referred to as PA_1 = (SB_1, LPA_1, $MVPA_0$).

5. With the modified observation PA_1, employ the same fitted PTR model to predict the outcome, denoted the predicted outcome as P_1.

Then the change in the predicted outcome, denoted as $\Delta P = P_1 - P_0$, can be considered as the predicted substitution effect between SB and LPA. This signifies how the predicted outcome would be affected by allocating the time spent in SB to LPA, specifically by decreasing δ minutes in SB and increasing δ minutes in LPA. When it comes to interpreting the ILR-transformation regression, we can employ a similar approach to examine the substitution effect among various pairs of PA categories.

Chapter 4

POWER TRANSFORMATION-BASED REGRESSION FOR COMPOSITIONAL DATA CONTAINING ZERO VALUES

Despite the applicability of PTR to compositional data that includes zero values, it is imperative to avoid setting the transformation parameter to zero or very small values. This is due to the inherent properties associated with the power transformation, as outlined in Section 2.3. This section aims to comprehensively address these concerns. We propose two estimation methods for PTR: constrained maximum likelihood estimation and modified likelihood procedures.

4.1 Issues Arising from Zero Values in Compositional Data Analysis

Compositional data analysis poses significant challenges due to the presence of zero values. The simplex defined in equation (1.1) allows for zero values in the data, but some modeling techniques are limited to compositional data without zero values. The presence of zero values is a major obstacle for methods that rely on the log-ratio quantity and requires special treatment. The origins of zero values in compositional data can be categorized into two distinct types: rounded zeros or values that fall below the limit of detection, and structural or essential zeros, which necessitate separate handling techniques.

Aitchison [2003] initially propose a zero-value replacement method prior to the application of log-ratio transformations. This approach assumes that zero values are likely due to imprecise measurement and represent rounded zeros. For instance in geology, some values may fall below the detection limit of the instrument used. In such cases, these values are usually considered missing, and missing value imputation techniques are applied. However, the unique properties of the simplex sample space require a different approach than that used in typical missing value imputation methods. Martín-Fernández [2003] compare various non-parametric imputation methods and propose a generalization of Aitchison's multiplicative approach (Aitchison [2003]). An al-

ternative approach is to substitute zero values with a very small quantity, as proposed by Zadora and Neocleous [2009]. By contrast, Palarea-Albaladejo et al. [2007] and Palarea-Albaladejo and Martín-Fernández [2008] propose parametric methods that utilize the EM algorithm after the logarithmic transformation of the data to replace rounded zeros. Hron et al. [2012] introduce the k-NN procedure using a logarithm-based distance metric and further propose an iterative model-based imputation technique that improves upon this procedure.

On the other hand, zero values in compositional data can be true and represent structural or essential zeros. For instance, when examining the time spent on the highest-intensity activity behavior from a less active population, the presence of zero values may reflect the absence of such activity rather than measurement error. Aitchison [2003] propose a conditional multivariate normal distribution for dealing with such structural zeros, which requires zeros to be in the same pattern. Since then, many other approaches have been proposed for handling essential zero values, either by treating them naturally or by using appropriate distributions inspired by Aitchison's work. One such approach is the square root transformation, which allows for the modeling of compositional data without modifying zero values (Scealy et al. [2015], Scealy and Welsh [2011], Scealy and Welsh [2014]). By contrast, Bear and Billheimer [2016] suggest a multivariate normal, while Aitken et al. [2007] adopt a zero-inflated distribution to deal with this problem. Another approach is the mixture model based on the multivariate skew-normal distribution developed by Stewart and Field [2011], which allows for greater flexibility compared to the multivariate normal.

Here we presented a toy example to illustrate the impact of different values of α on transformed values when dealing with compositional data that contained zero values. We set the sample size $N = 500$ with $D = 3$ and simulated data $x_i = (x_{i1}, x_{i2}, x_{i3})$ from the independent multivariate normal distribution $MVN_D(\mu, \Sigma)$, where $\mu = (10, 5, 1)$ and $\Sigma = \text{diag}(2, 0.5, 0.3)$, for $i = 1, 2, \ldots, N$. Additionally, binary indicators h_i were generated from a Bernoulli distribution with probability $p = 0.9$. Compositional data containing zero values were obtained through $x_i = (x_{i1}, x_{i2}, x_{i3} \cdot h_i)$. Figure 4.1 illustrated the transformed data considering various values of α, ranging from 0.005 to 1, where a small value of $\alpha = 0.005$ was chosen, while $\alpha = 1$ indicated that no transformation was applied. From the figure, it can be observed that when the data contained zero compositions, applying a power transformation with a small α (e.g., 0.005, 0.05) resulted in transformed data that deviated

significantly and exerted high leverage compared to the transformed data derived from non-zero compositional data. Consequently, this led to challenges in modeling fitting. However, when α was relatively large (e.g., 0.3, 0.7), these issues were alleviated, emphasizing the importance of avoiding small estimates of α. Furthermore, similar problems arise when we substituted the zero compositions with small values such as 0.001/1440. As shown in Figure 4.2, we can still observe similar patterns where the small compositions were defined as compositions that were smaller than 0.001/1440.

4.2 A Constrained Maximum Likelihood Approach

When it comes to modeling zero-contained compositional data, the model and estimation procedures discussed in Section 3 remain mostly valid. However, it is essential to note that the transformation parameter α must be restricted to the range of (0,1] in this context, and it cannot be set to zero. This is because when $\alpha \to 0$, the power transformation converges to the ILR-transformation, which is not well defined in the presence of zero values. Moreover, problems with model fitting and unexpected prediction performance may arise in the presence of zeros in compositional data when the value of α is close to zero. This is attributable to the behavior of the transformed values, $\mathbf{z}^{(\alpha)}$, defined in equation (2.2), which tend to approach infinity as α approaches zero. These issues may result in the estimates of coefficients being extremely small for these very large $\mathbf{z}^{(\alpha)}$, even if only a small subset of observations contain zero values. As a result, such problems may give rise to unrealistic model-fitting performances.

To avoid these issues resulting from small estimates of α, a straightforward approach is to introduce a constraint on α, such as enforcing it to fall within the interval $[c,1]$, where c is a constant satisfying the conditions $0 < c \leq 1$. This constraint serves the purpose of preventing the estimate of α from reaching zero during the implementation of the maximum likelihood technique. This leads to the formulation of the constrained maximum likelihood (ML) estimator of α, denoted as $\hat{\alpha}_c$, which is defined as follows:

$$\hat{\alpha}_c = \underset{\alpha \in [c,1]}{\operatorname{argmax}} \mathcal{L}(\alpha, \beta, \sigma^2), \tag{4.1}$$

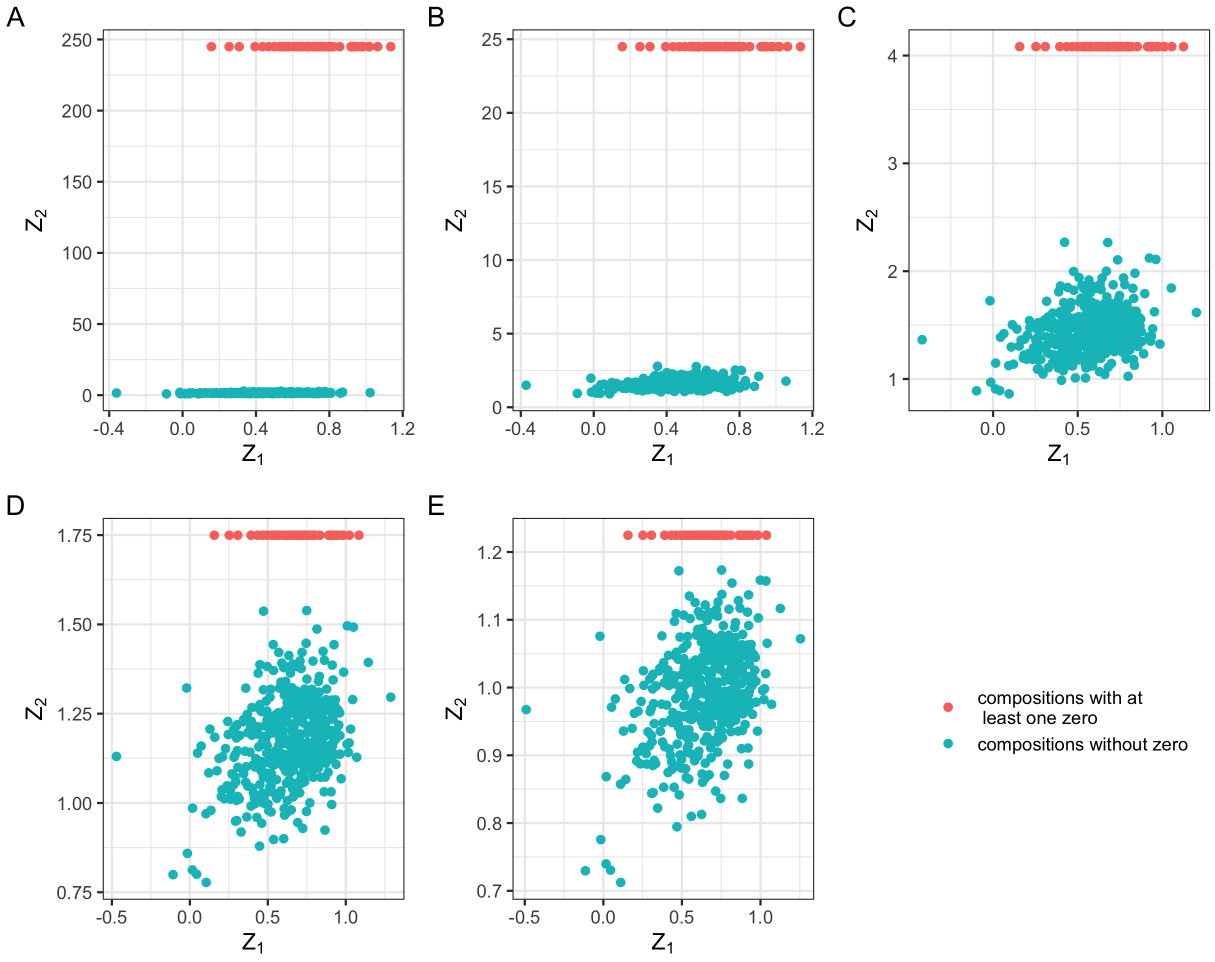

Figure 4.1: Different values of α and corresponding transformed data, which contain zero compositions: each figure shows the first column of the transformed data, $Z_1^{(\alpha)}$, versus the second column of the transformed data, $Z_2^{(\alpha)}$. A-E correspond to $\alpha = 0.005, 0.05, 0.3, 0.7, 1$.

where $\ell(\alpha, \beta, \sigma^2)$ represents the log-likelihood of the complete data as defined in Equation (3.6), and c represents the constraint.

Due to the introduction of the constraint c, the asymptotic properties of the constrained ML-based estimator, $\hat{\alpha}_c$, are different from what has been outlined in Section 3.2.

Proposition 6. *When α_0 is an interior point or on the boundary of its parameter space [0,1] and c is the constraint, the asymptotic distribution of $\hat{\alpha}_c$ has the following form:*

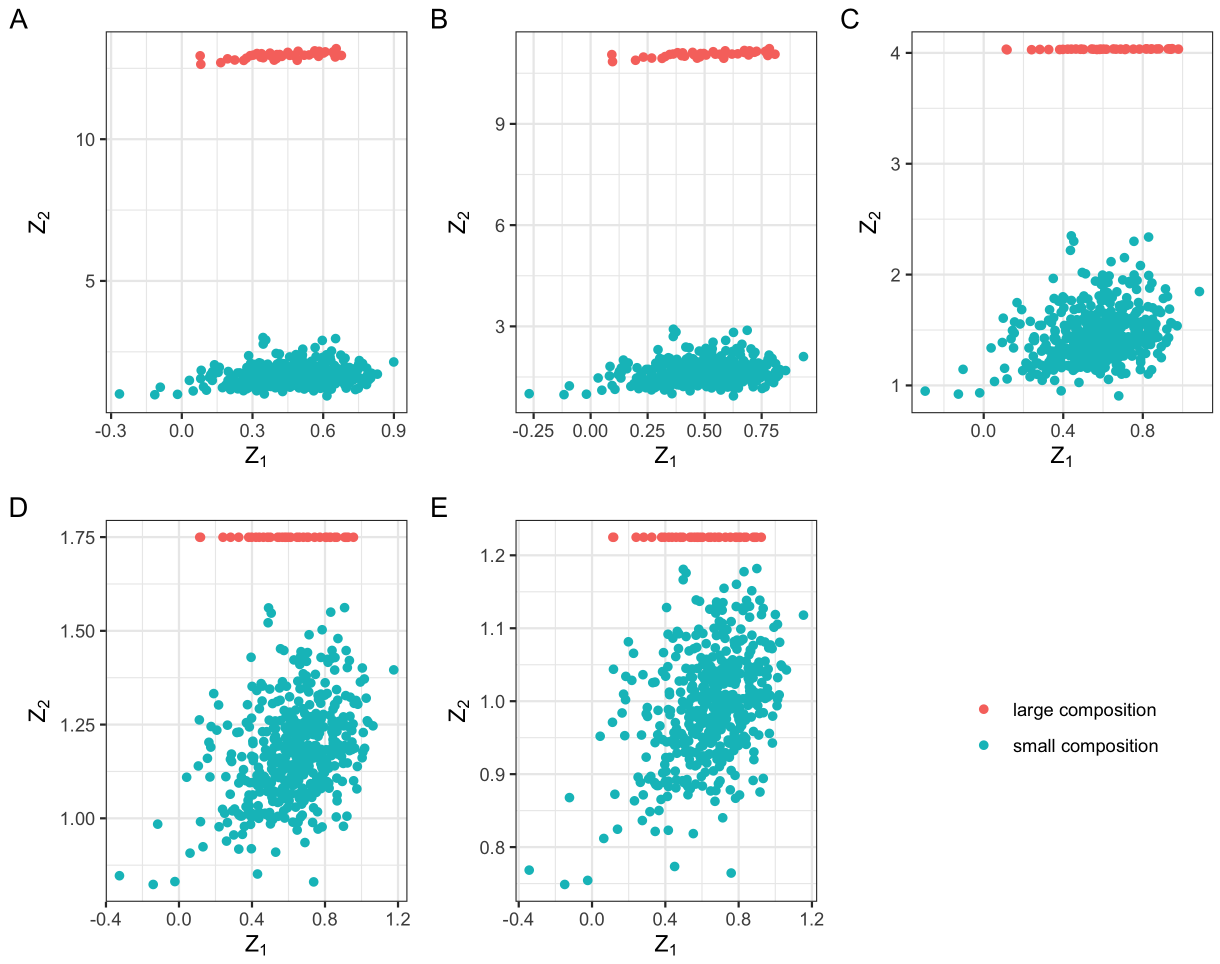

Figure 4.2: Different values of α and corresponding transformed data, whose zero compositions are replaced with small values: each figure shows the first column of the transformed data, $Z_1^{(\alpha)}$, versus the second column of the transformed data, $Z_2^{(\alpha)}$. Small compositions are defined as compositions that are smaller than 0.001/1440. A-E correspond to $\alpha = 0.005, 0.05, 0.3, 0.7, 1$.

- *if* $\alpha_0 \leq c$, $\sqrt{N}(\hat{\alpha}_c - \alpha_0) \xrightarrow{d} N^c(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$, *where* $N^c(0, [\boldsymbol{I}_n^{-1}]_{22})$ *represents* $N(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$ *being left truncated at the* $c - \alpha_0$.

- *if* $c < \alpha_0 < 1$, $\sqrt{N}(\hat{\alpha}_c - \alpha_0) \xrightarrow{d} N(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$.

- *if* $\alpha_0 = 1$, $\sqrt{N}(\hat{\alpha}_c - \alpha_0) \xrightarrow{d} N^-(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$, *where* $N^-(0, [\boldsymbol{I}_n^{-1}]_{22})$ *represents* $N(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$

being right truncated at 0.

In summary, the constrained ML-based estimator $\hat{\alpha}_c$ strikes a balance between regulating transformed covariates $\mathbf{z}^{(\alpha)}$ and maintaining prediction performance. When the true value of α exceeds the constraint c, the estimation performance of $\hat{\alpha}_c$ matches that of the ML-based $\hat{\alpha}$, allowing it to inherit the desirable properties of $\hat{\alpha}$. However, when the true value of α falls below the constraint c, implementing the constrained maximum likelihood approach effectively mitigates the challenges posed by largely transformed covariates at the expense of a slight decrease in prediction performance. This trade-off will be demonstrated in the simulation studies.

Hence, based on the preceding discussion, the selection of c is crucial for the successful implementation of such methods and requires careful consideration. A very small c fails to significantly alleviate the problems arising from large covariates $\mathbf{z}^{(\alpha)}$, particularly when the true value of α approaches zero. Conversely, a very large c leads to the excessive sacrifice of prediction performance, impeding the flexibility of the proposed PTR method, especially when the true value of α is close to zero. To determine an appropriate value for c, we conducted a series of simulations using different values of c ranging from very small to 1 and evaluated their prediction and corresponding transformed covariates $\mathbf{z}^{(\hat{\alpha}_c)}$. Further details on the simulation results will be discussed in Section 5.4.

4.3 A Modified Likelihood Approach

We also consider an alternative approach by introducing a modified likelihood function, inspired by the work of Chen et al. [2001]. Specifically, we incorporate a penalty term on $\log(\alpha)$ with the intention of discouraging α from being in proximity to zero. Specifically, we define the modified likelihood $\ell_m(\theta)$ as follows:

$$\mathcal{L}_m(\alpha,\beta,\sigma^2) = \mathcal{L}(\alpha,\beta,\sigma^2) + W\log(\alpha), \tag{4.2}$$

where $\mathcal{L}(\alpha,\beta,\sigma^2)$ is the log-likelihood function of the complete data as specified in (3.6), and W denotes the coefficient for the $\log(\alpha)$ penalty. Subsequently, we can obtain the estimator $\hat{\alpha}_m$ by

maximizing the modified likelihood $\mathcal{L}_m(\alpha, \beta, \sigma^2)$:

$$\hat{\alpha}_m = \underset{\alpha \in (0,1]}{\operatorname{argmax}} \mathcal{L}_m(\alpha, \beta, \sigma^2). \tag{4.3}$$

It is worth noting that the asymptotic properties discussed in Section 3.2 still hold for the ML-estimator of the modified likelihood, $\hat{\alpha}_m$, even when a penalty term is incorporated. This is because, in the limit as the sample size N approaches infinity, the likelihood function $\ell(\alpha)$ becomes dominant over the modified likelihood. Consequently, the modified-likelihood-based estimator demonstrates identical asymptotic properties to the likelihood-based estimators.

Proposition 7. *If α_0 is an interior point of the interval [0,1], the asymptotic distribution of $\hat{\alpha}_m$ can be expressed as*

$$\sqrt{N}(\hat{\alpha}_m - \alpha_0) \xrightarrow{d} N(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22}). \tag{4.4}$$

Proposition 8. *When $\alpha_0 = 1$, which is on the boundary of its parameter space, the asymptotic distribution of $\hat{\alpha}_m$ has the following form:*

$$\sqrt{N}(\hat{\alpha}_m - \alpha_0) \xrightarrow{d} N^{-}(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22}), \tag{4.5}$$

where $N^{-}(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$ represents $N(0, [\boldsymbol{I}(\theta_0)^{-1}]_{22})$ being right truncated at 0.

The modified ML-based estimator, $\hat{\alpha}_m$, also achieves a balance between the regularization of transformed covariates $\mathbf{z}^{(\alpha)}$ and the prediction performance. However, it adopts a different approach compared to the constrained-based estimator. By introducing a penalty term W on α, the estimate of α is discouraged from approaching zero. Consequently, this results in larger estimates of α compared to the original ML-based estimator, $\hat{\alpha}$. As a result, the issues arising from very small estimates of α can be alleviated. On the on hand, a small value of W fails to significantly mitigate the issues arising from large covariates $\mathbf{z}^{(\alpha)}$, as it exerts a negligible penalty on α and consequently still produces small estimates of α, particularly when the true value of α approaches zero. On the other hand, an excessively large W imposes an excessive penalty on α, yielding a substantially larger estimate of α. This, in turn, leads to an overemphasis on the penalty term at the expense of prediction performance, particularly when the true value of α is close to zero.

The inherent trade-off between regularization and prediction performance will be demonstrated in simulation studies.

Therefore, the selection of W holds paramount importance for the successful implementation of such methods and necessitates meticulous consideration. To ascertain suitable values for W, extensive simulation studies were conducted, encompassing a wide range of different W. These studies allowed for a thorough evaluation of their prediction performances, as well as the corresponding transformed covariates, $\mathbf{z}^{(\hat{\alpha}_c)}$. Consequently, recommendations were formulated regarding the choice of W across various sample sizes based on the simulation settings. Comprehensive details regarding the simulation results and the underlying rationale behind these recommendations will be extensively discussed in the subsequent simulation sections.

Remark: Here, we present an alternative interpretation that establishes a connection between the proposed constrained maximum likelihood approach and the modified likelihood approach. It is evident that the modified likelihood approach incorporates a penalty-based method, which is also utilized in the constrained maximum likelihood approach. In the case of the constrained approach, once the constraint c is determined, the estimated value $\hat{\alpha}_c$ cannot be smaller than c, while any estimate equal to or greater than c remains unaffected. This implies that the constrained method imposes an infinite penalty on $\hat{\alpha}_c$ when it is smaller than c, while it imposes no penalty when $\hat{\alpha}_c$ is equal to or greater than c.

Chapter 5

SIMUALTION STUDIES

In this section, our attention was directed toward the simulation studies conducted on the power transformation-based approaches proposed in the preceding chapters. The aim of these studies was to explore the finite sample properties of the methods described in Chapters 3 and 4 across three distinct scenarios.

5.1 Simulation Settings

We set the true value of α to be 0.05, 0.2, 0.5, 0.7, 0.95, 1, and the true value of $\beta = (\beta_0, \beta_1, \beta_2)$ to be $[0.2, 5, 3]^T$ with the number of dimensions $D = 3$. The sample size was set to be $N = 100, 200, 500, 1000, 2000$, and the number of simulation replications $M = 1000$. We considered two simulation scenarios for generating compositional data. In **Scenario I**, we simulated data $x_i = (x_{i1}, x_{i2}, x_{i3})$ from a multivariate normal distribution $MVN_D(\mu, \Sigma)$ independently, where $\mu = (10, 5, 1)$ and $\Sigma = \text{diag}(2, 0.5, 0.3)$, $i = 1, 2, \ldots, N$. In **Scenario II**, we first generated x_i in the same way as **Scenario I** and binary indicators h_i from a Bernoulli distribution with probability $p = 0.7$. Compositional data containing zero values were obtained by applying a normalization technique. The normalization formula used was as follows: $x_i = (|x_{i1}|, |x_{i2}|, |x_{i3}| \cdot h_i)/(|x_{i1}| + |x_{i2}| + |x_{i3}| \cdot h_i)$. This approach guaranteed that the compositional data remained non-negative and summed up to 1. Next, the power transformation with specific α was applied to x_i to obtain $z_i^{(\alpha)} = (z_{i1}^{(\alpha)}, z_{i2}^{(\alpha)})$. Then the outcome $\mathbf{Y}$ was generated using $\mathbf{Y} = (1, [z_1^{(\alpha)}, z_2^{(\alpha)}])^T \beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma_N^2)$, with σ_N chosen to allow the signal-to-noise ratio (SNR) to vary among three values $0.5, 1, 2$.

Given these parameters, the observed data were generated independently from corresponding distributions. However, due to the definition of α, certain constraints must be taken into account when estimating $\hat{\alpha}$. In our simulations, we explored various techniques for incorporating this restriction. For the implementation of the GCV method, we assigned grid values to α

as $[\exp(a.grid), 0.001, 0.002, \ldots, 0.999, 1]$, where $a.grid$ consisted of 50 evenly spaced points between -20 and -7. In the case of MLE, we restricted the value of α to the interval $[1 \times 10^{-10}, 1]$ and utilized the L-BFGS-B algorithm (Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints) as presented in Byrd et al. [1995], which allows for the inclusion of simple box constraints on variables while maximizing the likelihood.

To further investigate the performances of the constrained and modified likelihood approach, we conducted additional simulations, referred to as **Scenario III**. This scenario retained most of the settings from **Scenario II** but incorporated several key modifications. Specifically, we varied the true value of α to be $0.005, 0.05, 0.3, 0.7$. The constraint c specified in (4.1) took values ranging from 1×10^{-10} to 1. We set the values of W defined in (4.3) from the exponential of $w.grid$, where $w.grid$ consisted of 1000 evenly spaced points between -10 and 15. During the implementation of the modified likelihood approach, the values of α were confined to the interval $[1 \times 10^{-10}, 1]$. Furthermore, we maintained a constant value of $\sigma_N = 5$ and set the probability p in the Bernoulli distribution to be 0.9.

5.2 Results under Simulation Scenario I

We found that finite sample performances of the ML and GCV-based estimators were similar under **Scenario I**. As the sample size increased, the variance and bias of $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$ decreased, as shown in Figure 5.1. It was noteworthy that the distributions of $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$ were asymmetric especially when the true α was near the boundaries of its parameter space. This phenomenon was particularly evident when the true α was exactly at the boundary of its parameter space (i.e., $\alpha = 0$ or 1) when the ML-based estimators asymptotically followed a truncated normal distribution.

The estimation of α had a moderate impact on the estimation of β. As such, we investigated the relationship between the bias of $\hat{\alpha}$ and $\hat{\beta}$. We found that when the true α was near or at the boundary of its parameter space, $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ exhibited systematic patterns in terms of their bias. For instance, when the true value of α, denoted as α_t, was 0.05, Bias($\hat{\alpha}$) and Bias($\hat{\beta}_1$) were positive, while Bias($\hat{\beta}_2$) was negative (Figure 5.1, Figure 5.2 [**C-H**]). The presence of a positive bias in $\hat{\alpha}$ resulted in a situation where $z_1^{(\hat{\alpha})}$ being larger than $z_1^{(\hat{\alpha}_t)}$. Consequently, during the regression analysis utilizing $z_1^{(\hat{\alpha})}$, an underestimation of β_1 occurred in an attempt to counterbalance

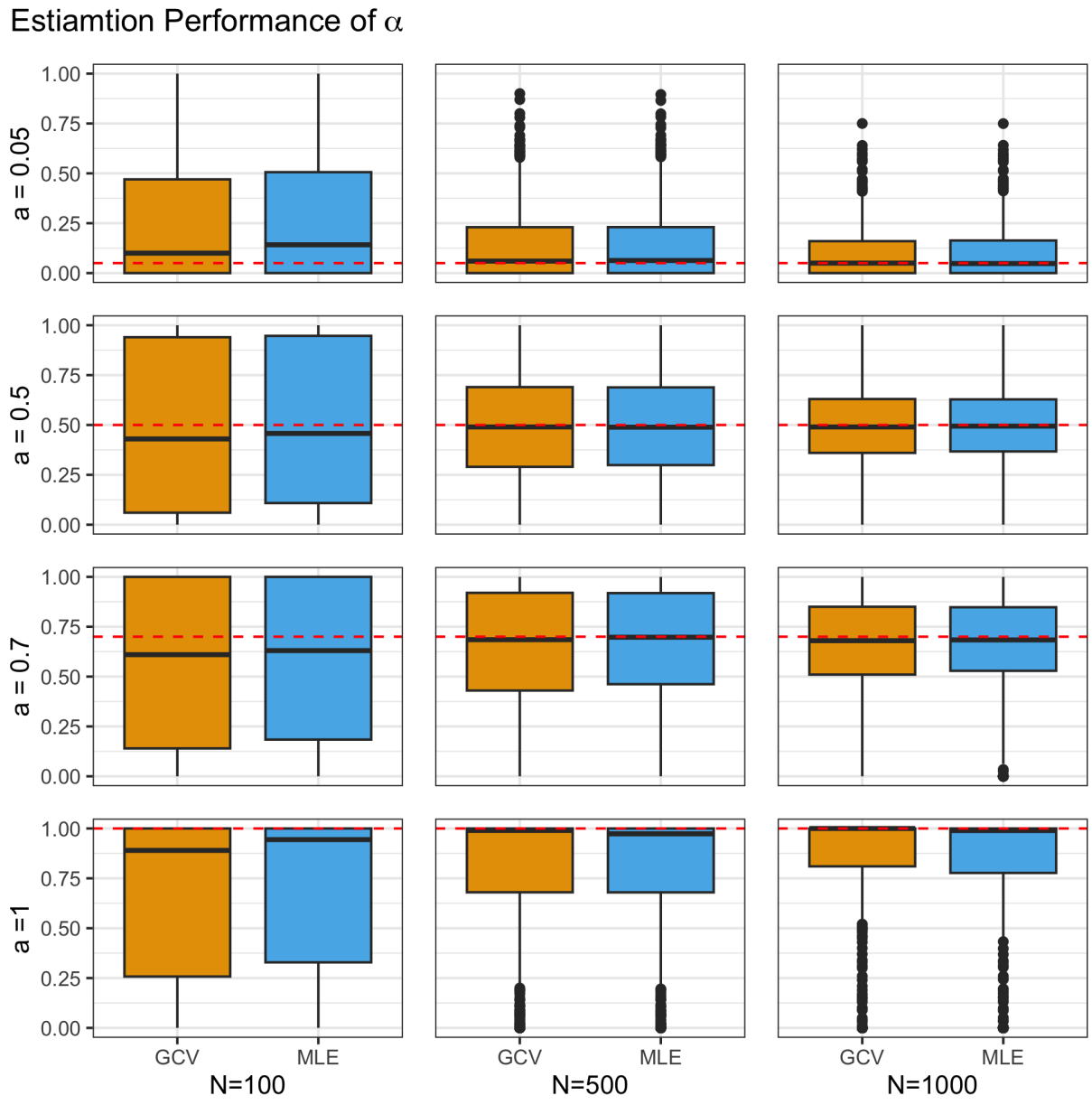

Figure 5.1: Estimation performance of $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$ under **Scenario I**: true α takes the value of 0.05, 0.5, 0.7, 1, sample size $N = 100, 500, 1000$. Each box plot has a unique setting with certain true α and N. For those plots, the sample size increases from left to right, and true α increases from top to bottom.

the influence of $z_1^{(\hat{\alpha})}$, thereby introducing systematic biases into $\hat{\beta}_1$. Likewise, the positive bias exhibited by $\hat{\alpha}$ led to a scenario where $z_2^{(\hat{\alpha})}$ was smaller than $z_2^{(\hat{\alpha}_t)}$. Hence, when performing the regression analysis with $z_2^{(\hat{\alpha})}$, an overestimation of β_2 transpired as an attempt to compensate for the impact of $z_2^{(\hat{\alpha})}$, thereby introducing systematic biases into $\hat{\beta}_2$. Moreover, when α_t assumed a large value close to 1, the negative bias pattern observed in $\hat{\alpha}$ also affected the biases of $\hat{\beta}_1$ and $\hat{\beta}_2$, manifesting as positive and negative biases, respectively.

Standard errors of $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$ were shown in Figure 5.2 [**A-B**]. The empirical standard errors of $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$ were similar. The difference between the model-based and empirical standard errors of $\hat{\alpha}$ decreased as the sample size and SNR increased. We also assessed predictive performances of the two methods by calculating their root mean squared prediction errors and observed similar performances (Table 5.1), which was not unexpected as the accuracy of predictions is largely determined by parameter estimation. When simulation errors were taken into consideration, the coverage probability of the 95% bootstrap confidence intervals for predicted outcomes varied in an acceptable range. However, neither MLE nor GCV showed a clear advantage over the other due to their similar model-fitting results (Table 5.2).

We conducted simulations to evaluate the performance of hypothesis testing mentioned in Section 3.3, with a focus on testing whether α lies on the boundary of its parameter space ($\alpha_0 = 0$ or 1). Using a significance level of 0.05 and the SNR set at 1, we varied the sample size N to be 100, 500, and 1000, and performed 1000 simulation runs. To assess the estimated proportions of rejecting the null hypothesis when testing α as indicated in (3.15), we tested $H_0 : \alpha = 0$ and $H_0 : \alpha = 1$ with $\alpha = 0.05$, 0.1, 0.3, 0.5, and 1 under alternatives. The simulation results were illustrated in Figure 5.5 (**A-B**). It clearly demonstrated that as the sample size increased, the Type I error converged to the expected Type I error rate of 0.05. Moreover, the power of the test increased with larger sample sizes and also rose as the discrepancy between α_t and its boundary value (0 or 1) becomes more pronounced.

5.3 Results under Simulation Scenario II

In **Scenario II**, the finite sample properties of $\hat{\alpha}$ and $\hat{\beta}$ were similar when α_t took relatively large values. However, when α_t was small and close to zero, the estimation performance of GCV was

Table 5.1: Prediction performances (measured as root mean squared prediction error) of fitted α-transformation regression model with GCV and MLE under **Scenario I** and **Scenario II**: true α takes the value of 0.05, 0.5, 0.7, 1, sample size $N = 100, 500, 1000, 2000$.

	α	Method	N=100	N=500	N=1000	N=2000
Scenario I	0.05	GCV	1.396	1.380	1.374	1.369
		MLE	1.397	1.380	1.374	1.369
	0.5	GCV	1.335	1.305	1.309	1.304
		MLE	1.335	1.305	1.309	1.304
	0.7	GCV	1.290	1.257	1.256	1.254
		MLE	1.290	1.257	1.256	1.254
	1	GCV	1.206	1.179	1.171	1.173
		MLE	1.206	1.179	1.171	1.173
Scenario II	0.05	GCV	33.089	32.314	32.124	32.061
		MLE	32.408	31.995	32.058	32.060
	0.5	GCV	2.297	2.236	2.224	2.223
		MLE	2.298	2.239	2.230	2.224
	0.7	GCV	1.592	1.575	1.567	1.566
		MLE	1.614	1.571	1.565	1.565
	1	GCV	1.267	1.237	1.234	1.230
		MLE	1.261	1.231	1.232	1.230

Table 5.2: Coverage probability of the bootstrap-based 95% confidence interval for predicted outcomes using PTR with GCV and MLE under **Scenario I** and **Scenario II**: true α takes the value of 0.05, 0.5, 0.7, 1, sample size $N = 100, 500, 1000, 2000$.

	α	Method	N=100	N=500	N=1000	N=2000
Scenario I	0.05	GCV	0.951	0.945	0.938	0.946
		MLE	0.922	0.940	0.942	0.940
	0.5	GCV	0.946	0.954	0.958	0.952
		MLE	0.946	0.960	0.958	0.947
	0.7	GCV	0.950	0.946	0.941	0.948
		MLE	0.939	0.939	0.947	0.942
	1	GCV	0.961	0.949	0.940	0.944
		MLE	0.946	0.937	0.947	0.939
Scenario II	0.05	GCV	0.963	0.955	0.939	0.950
		MLE	0.954	0.948	0.944	0.953
	0.5	GCV	0.952	0.938	0.935	0.950
		MLE	0.947	0.947	0.950	0.935
	0.7	GCV	0.957	0.965	0.962	0.962
		MLE	0.948	0.961	0.960	0.963
	1	GCV	0.957	0.973	0.949	0.956
		MLE	0.959	0.951	0.956	0.957

much worse than that of MLE. This difference can be attributed to the fact that when α_t was small, the empirical and model-based standard error of $\hat{\alpha}$ were similar, while the empirical-based standard error of $\hat{\alpha}_{GCV}$ was much larger than that of $\hat{\alpha}$ (Figure5.4 **D**). Figure5.4 [**A-C**] also illustrated how the large variance of $\hat{\alpha}_{GCV}$ contributed to its poor estimation performance. In order to comprehend the irregular behavior exhibited by GCV-based estimators, we conducted supplementary simulations with specific parameter values: $\alpha_t = 0.05$, $N = 100$, and an SNR of 0.5. The results, illustrated in Figure 5.3, revealed an intriguing observation: the maximum of MPSE could occur at any point within the interval $(0, 1]$, which consequently led to a substantial variance in the estimated value of $\hat{\alpha}_{GCV}$. Conversely, the ML-based estimator, $\hat{\alpha}$, were usually in the proximity of zero, specifically in the vicinity of α_t. This characteristic ensured a relatively smaller variance. Consequently, the prediction errors depicted in Figure 5.4 (**E**) and Table 5.1 remained similar for both MLE and GCV, with the exception of the scenario where $\alpha_t = 0.05$. Moreover, the simulation results for the coverage probability, presented in Table 5.2, demonstrated comparable behaviors between **Scenario I** and **Scenario II** for both estimation methods.

To assess the performance of the test described in (3.16), we conducted simulations to test whether α is on the boundary of its parameter space. We evaluated its estimated proportions of rejecting the null hypothesis $H_0 : \alpha = 1$, with $\alpha = 0.95, 0.9, 0.7$, and 0.5 under alternatives. Using a pre-specified significance level of 0.05, we found that the Type I error converged to the Type I error rate (0.05) as expected. The results depicted in Figure 5.5 (**C**) demonstrated that the power of the test increased as the sample size and the discrepancy between the true α and $\alpha_0 = 1$ increased.

5.4 Results under Simulation Scenario III

We first focused on discussing the results of implementing the constrained maximum likelihood approach. Our aim was to investigate the behavior of this estimation method with different constraints c and the corresponding prediction performance to inform optimal choices of c. For each $\hat{\alpha}_c$, we calculated its prediction error using the root mean squared prediction error (RMSPE). To ensure that the results were comparable for different α_t, we computed a scaled version of RMSPE. We represented the RMSPE of a specific c and α_t as RMSPE(c, α_t), and defined RMSPE$_{\max}$ as the maximum of RMSPE(c, α_t) over all possible values of c and α_t. With this, we can obtain the scaled

prediction error as the ratio of the logarithm of $\text{RMSPE}(c, \alpha_t)$ to the logarithm of $\text{RMSPE}_{\max}$

$$\widetilde{\text{RMSPE}}(c, \alpha_t) = \frac{\log[\text{RMSPE}(c, \alpha_t)]}{\log[\text{RMSPE}_{\max}]}. \tag{5.1}$$

We also introduced a metric for evaluating the performance of the transformed covariates $\mathbf{z}^{(\hat{\alpha}_c)}$. In this context, $\mathbf{z}_0^{(\alpha)}$ represents the transformed covariates obtained from compositional data containing zeros, whereas $\mathbf{z}_1^{(\alpha)}$ corresponds to the transformed covariates derived from non-zero compositional data. By calculating the mean of $\mathbf{z}_0^{(\hat{\alpha}_c)}$ and $\mathbf{z}_1^{(\hat{\alpha}_c)}$, denoted as $\mu_0(c)$ and $\mu_1(c)$ respectively, we can define the distance ratio metric as follows:

$$R(c) = \frac{||\mu_0(c) - \mu_1(c)||}{||\mu_0(c=1) - \mu_1(c=1)||}, \tag{5.2}$$

where $||\cdot||$ represents the L_2 norm. In the case of compositional data containing zeros, a small α significantly magnifies the transformed data, while a large α amplifies it to a lesser degree. Consequently, the distance between $\mu_0(c)$ and $\mu_1(c)$ can be utilized as a metric to quantify the influence of α on the transformation of data. As α approaches zero, the distance $||\mu_0(c) - \mu_1(c)||$ increases due to the increase in $\mu_0(c)$. To ensure the comparability of results, we selected $\mathbf{z}^{(\hat{\alpha}_{c=1})}$ as the reference, which is equivalent to $\mathbf{z}^{(\alpha=1)}$, indicating that no transformation is applied.

Figure 5.6 illustrated the prediction performances and the behaviors of the introduced distance ratio metric $R(c)$. When α_t exceeded the constraint c, the prediction error was minimally affected or remains unaffected by the constraint. However, if the constraint surpassed α_t, the error increased along with the constraint. This can be attributed to the widening gap between α_t and $\hat{\alpha}_c$ as c exceeded α_t. A similar trend was observed regarding the ratio metric. When c was smaller than α_t, $R(c)$ remained constant until c surpasses α_t, after which it dropped along with the increasing constraint. This phenomenon became particularly pronounced when α_t was small, such as $\alpha_t = 0.005$. However, for large values of α_t, such as 0.05, 0.3, and 0.7, the prediction error showed no significant changes. Given our primary objective of avoiding excessively small estimates of α, especially when α_t was indeed small, it was advisable to focus on the case where $\alpha_t = 0.005$. Taking into account the trade-off between prediction performance and distance metric depicted in Figure 5.6, we recommended choosing $c = 0.05$ as it resulted in an acceptable sacrifice in prediction without being overly substantial and improves the performance of the transformed covariates by signifi-

cantly reducing the ratio metric. However, the choice of the constraint c should depend on the specific context and structure of the data. The recommendations provided here were based solely on the designed simulation settings and should only be considered general guidelines.

Next, our focus shifted towards examining the performance of implementing the modified likelihood approach with respect to different values of W. For each W, we applied the modified likelihood approach to obtain $\hat{\alpha}_m$, denote as $\hat{\alpha}_{m,W}$, and also computed a scaled version of RMSPE as follows. We represented the RMSPE of a specific W and α_t as $\text{RMSPE}(W, \alpha_t)$, and defined $\text{RMSPE}_{\max}$ as the maximum of $\text{RMSPE}(W, \alpha_t)$ over all possible values of W and α_t. With this, we can obtain the scaled prediction error as the ratio of the logarithm of $\text{RMSPE}(W, \alpha_t)$ to the logarithm of $\text{RMSPE}_{\max}$

$$\widetilde{\text{RMSPE}}(W, \alpha_t) = \frac{\log[\text{RMSPE}(W, \alpha_t)]}{\log[\text{RMSPE}_{\max}]}. \tag{5.3}$$

We utilized a similar metric for evaluating the effect of $\hat{\alpha}_{m,W}$ on the transformation of data, denoted as $\mathbf{z}^{(\hat{\alpha}_{m,W})}$. By calculating the respective means of $\mathbf{z}_0^{(\hat{\alpha}_{m,W})}$ and $\mathbf{z}_1^{(\hat{\alpha}_{m,W})}$, denoted as $\mu_0(W)$ and $\mu_1(W)$, we can define the distance ratio metric as follows:

$$R(W) = \frac{||\mu_0(W) - \mu_1(W)||}{||\mu_0 - \mu_1||}, \tag{5.4}$$

where μ_0 and μ_1 denote the mean of $\mathbf{z}_0^{(\alpha=1)}$ and $\mathbf{z}_1^{(\alpha=1)}$, respectively.

Figure 5.7, 5.8, 5.9, and 5.10 depicted the prediction performances and the behavior of the introduced distance ratio metric $R(W)$ for sample sizes $N = 100, 500, 1000, 2000$, respectively. These metrics exhibited similar patterns across different sample sizes. When W was relatively small, the prediction error shows minimal or no significant impact. This occurred because, during this phase, the penalty on α being close to zero is small and the original likelihood still dominates the modified likelihood. As a result, the modified-ML-based estimator $\hat{\alpha}_m$ closely approximated the ML-based estimator $\hat{\alpha}$, leading to similar prediction performances of the fitted models. However, as W increased, indicating a larger penalty on α, the modified-ML-based estimator $\hat{\alpha}_m$ tended to yield larger estimates for α in order to prevent the estimate from approaching zero, leading to the increase in the prediction error as W became larger. A similar trend was observed for the ratio metric. When W was relatively small, $R(W)$ remained almost unchanged, but it subsequently dropped as

W increased.

Similarly, this phenomenon became particularly pronounced when α_t was small, such as $\alpha_t = 0.005$, and when α_t took large values, such as 0.05, 0.3, and 0.7, the prediction error remained relatively unchanged. By considering the tread-off between the prediction performance and the distance metric as illustrated in Figures 5.7, 5.8, 5.9, and 5.10, we have formulated the following recommended range for W based on different sample sizes: For a sample size of $N = 100$, W should fall within the range of $[20, 60]$; for $N = 500$, the recommended range is $[100, 300]$; for $N = 1000$, the range is $[200, 600]$, and for $N = 1200$, it is $[400, 1200]$. All these suggested values of W struck a suitable balance, ensuring acceptable prediction outcomes without being excessively large, while simultaneously alleviating the effect of $\hat{\alpha}_m$ on the transformation of data by significantly reducing the ratio metric. Again, it was important to note that the choice of W should depend on the specific context and structure of the data. The recommendations provided here are based solely on the designed simulation settings and should only be considered as general guidelines.

We also applied four different methods that have been mentioned before to compare their prediction performances: ISM, ILR-transformed regression, the proposed constrained maximum likelihood approach ($c = 0.05$) of PTR, and the proposed modified likelihood approach of PTR ($W = 20$ for $N = 100$, $W = 100$ for $N = 500$, $W = 200$ for $N = 1000$, $W = 400$ for $N = 2000$).

In order to implement the ILR-transformed regression for compositional data containing zeros, we replaced zeros with a small value of 0.5 minutes. The RMSPE results obtained from these methods are presented in Table 5.3. Notably, when the values of α_t were relatively small, specifically $\alpha_t = 0.005$ and $\alpha_t = 0.05$, the ISM exhibited the highest RMSPE. However, the ILR-transformed regression produced a smaller, albeit still significantly larger RMSPE compared to the constrained and modified likelihood approach. Nevertheless, as the sample size increases for large α_t, the RMSPE values of all four approaches tend to be similar.

5.5 Summary

In this section, we have presented the simulation results of the PTR model under three different scenarios. We initially evaluated two estimation methods: maximum likelihood estimation and generalized cross-validation, and compared their performance when analyzing data with and with-

Table 5.3: Prediction performances (measured as root mean squared prediction error) of different methods under **Scenario III**: isotemporal substitution model, ILR-transformation regression (replace 0 with small values, 0.5 minutes), PTR with constrained estimation ($c = 0.05$) and modified estimation ($W = 20$ for $N = 100$, $W = 100$ for $N = 500$, $W = 200$ for $N = 1000$, $W = 400$ for $N = 2000$). True α takes the value of 0.05, 0.5, 0.7, 1, sample size $N = 100, 500, 1000, 2000$. *Abbrevation: Isotemporal substitution model, ISM. ILR-transformation regression, ILR. PTR with constrained maximum likelihood approach, PTR-C. PTR with modified likelihood approach, PTR-M.

	Method	N=100	N=500	N=1000	N=2000
$\alpha = 0.005$	ISM	145.958	149.709	149.944	150.143
	ILR	70.218	71.807	72.040	72.111
	PTR-C	17.215	18.753	18.990	19.284
	PTR-M	5.424	5.488	5.564	5.590
$\alpha = 0.05$	ISM	14.323	14.610	14.662	14.685
	ILR	7.769	7.895	7.922	7.935
	PTR-C	4.917	4.981	4.994	4.993
	PTR-M	5.322	5.369	5.381	5.376
$\alpha = 0.3$	ISM	5.080	5.152	5.163	5.167
	ILR	4.919	4.985	4.995	5.000
	PTR-C	4.881	4.983	4.983	4.990
	PTR-M	5.057	5.141	5.163	5.171
$\alpha = 0.7$	ISM	4.920	4.992	4.994	4.997
	ILR	4.920	4.994	4.996	4.998
	PTR-C	4.893	4.986	4.989	4.998
	PTR-M	4.924	4.988	4.994	4.995

out zero compositions. Under **Scenario I**, the results demonstrated that both MLE and GCV were similar in performance. Therefore, we concluded that both MLE and GCV are suitable when there are no zeros in the dataset. However, when the true value of α is relatively small under **Scenario II**, ML-based estimators showed better performance compared to GCV-based estimators due to the smaller standard errors. As it is often difficult to determine the true value of α in real data analysis, the ML-based estimator is recommended for its stable performance in simulation studies. Therefore, for the purpose of accurately estimating α and predicting outcomes, the ML-based estimator should be preferred, especially when the true α is uncertain or expected to be small.

Under **Scenario III**, we investigated the behavior of the proposed constrained maximum likelihood approach using different constraint values c, and modified likelihood approach with varying the penalty coefficient W. Based on the simulation results, we made recommendations regarding the choice of c and W. Additionally, we applied the ISM, ILR-transformed regression, and compared their prediction performance to the proposed methods. Notably, our findings demonstrated that the proposed two estimation procedures outperform the ISM and ILR-transformed model, particularly when the value of α_t is small.

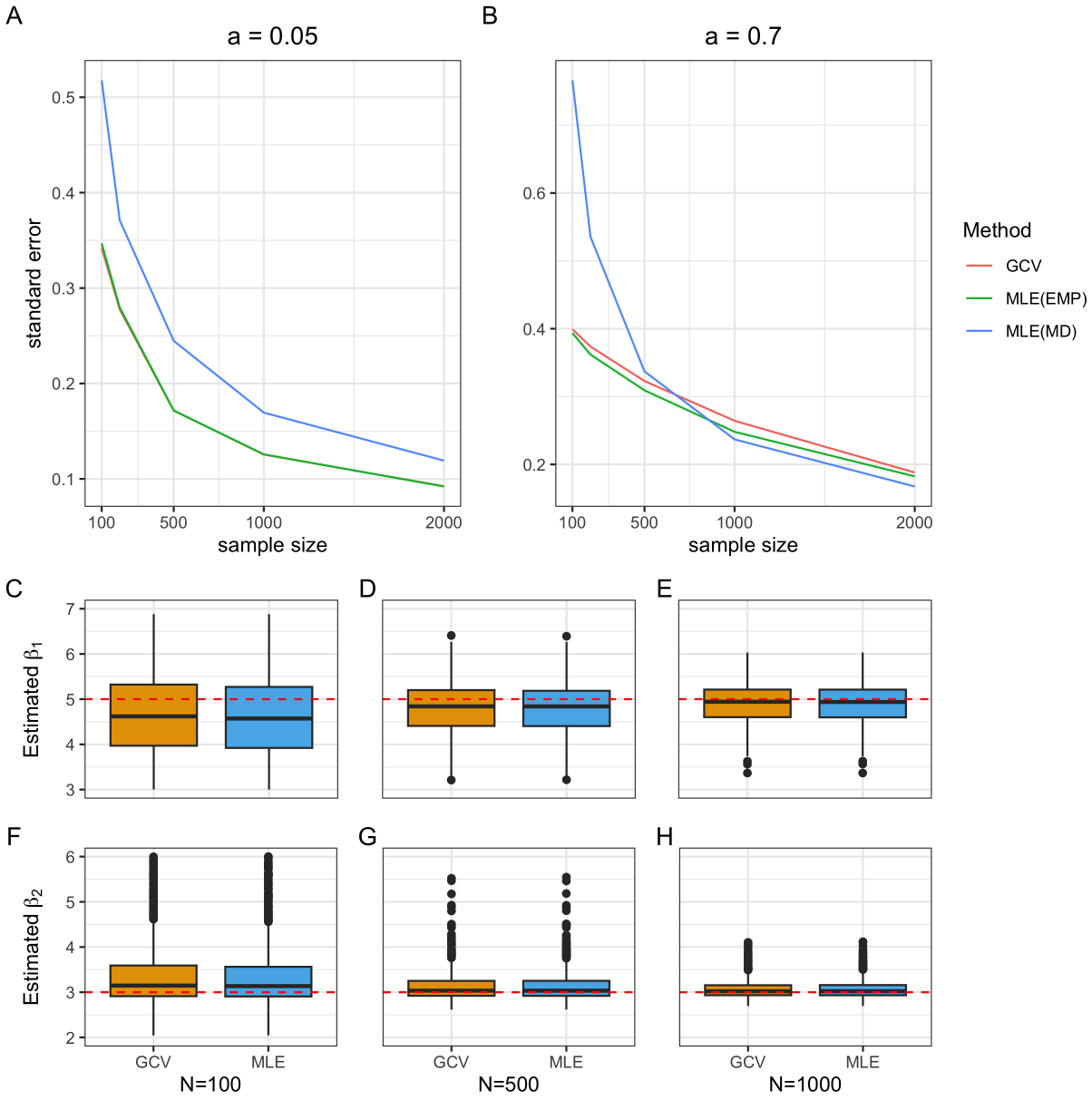

Figure 5.2: Simulation results under **Scenario I**. **A-B**: Standard errors of $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$, true $\alpha = 0.05, 0.7$. The empirical-based standard error of GCV, and both the empirical and model-based standard error of MLE were considered. **C-E**: Estimation performances of $\hat{\beta}_1$ with MLE and GCV, true $\alpha = 0.05$, sample size $N = 100, 500, 1000$. Each box plot has a unique setting with a certain N. For those plots, the sample size increases from left to right. **F-H**: Estimation performances of $\hat{\beta}_2$ with MLE and GCV, true $\alpha = 0.05$, sample size $N = 100, 500, 1000$. Each box plot has a unique setting with a certain N. For those plots, the sample size increases from left to right. *Abbrevation: Empirical, EMP. Model-based, MD.

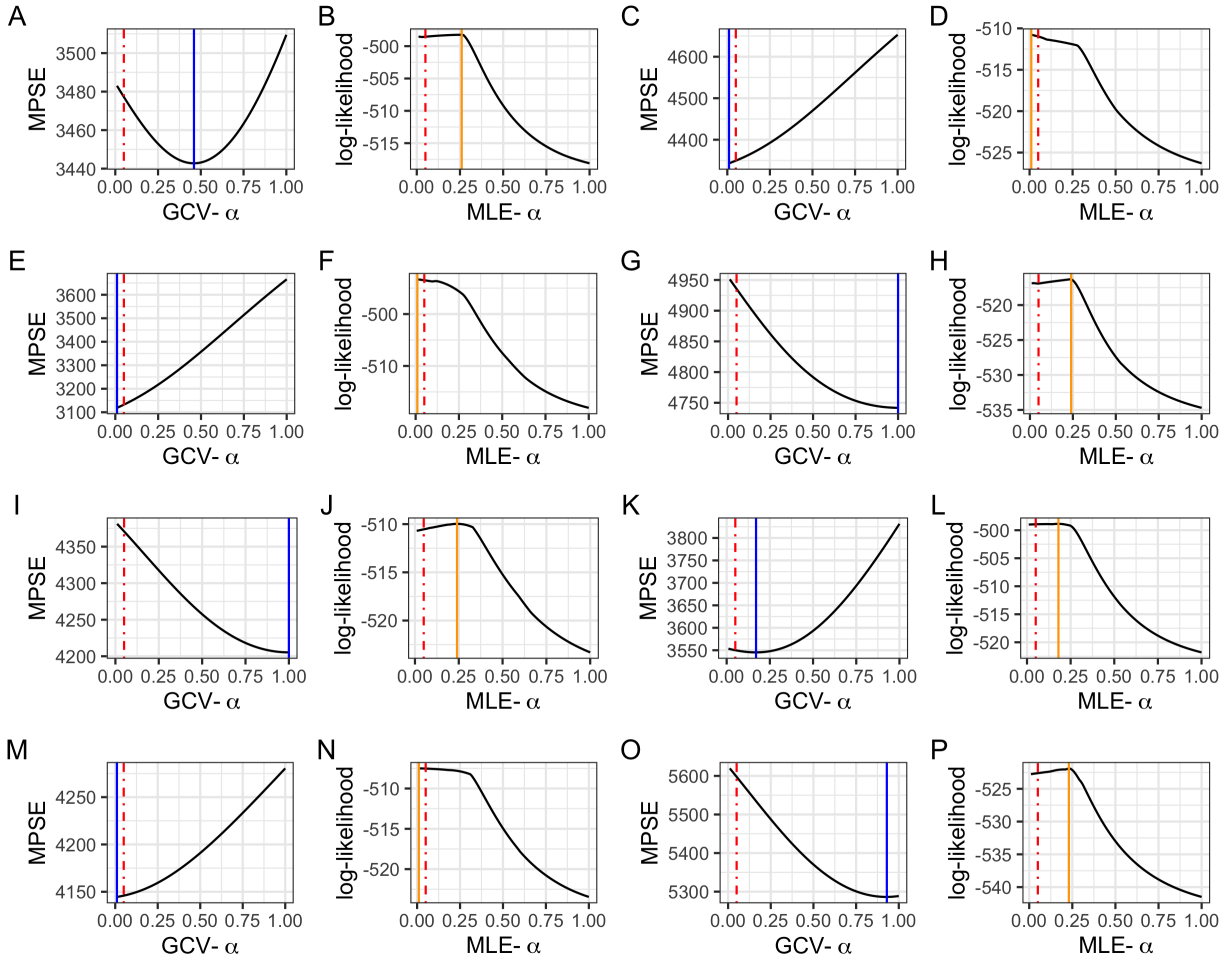

Figure 5.3: The behavior of the Mean Predicted Square Error (MPSE) and log-likelihood when estimating α with GCV and MLE, respectively. Simulations were conducted under **Scenario II**: true $\alpha = 0.05$, $N = 100$. AB, CD, EF, GH, IJ, KL, MN, and OP are pairs of results of MPSE and log-likelihood for 8 simulated datasets.

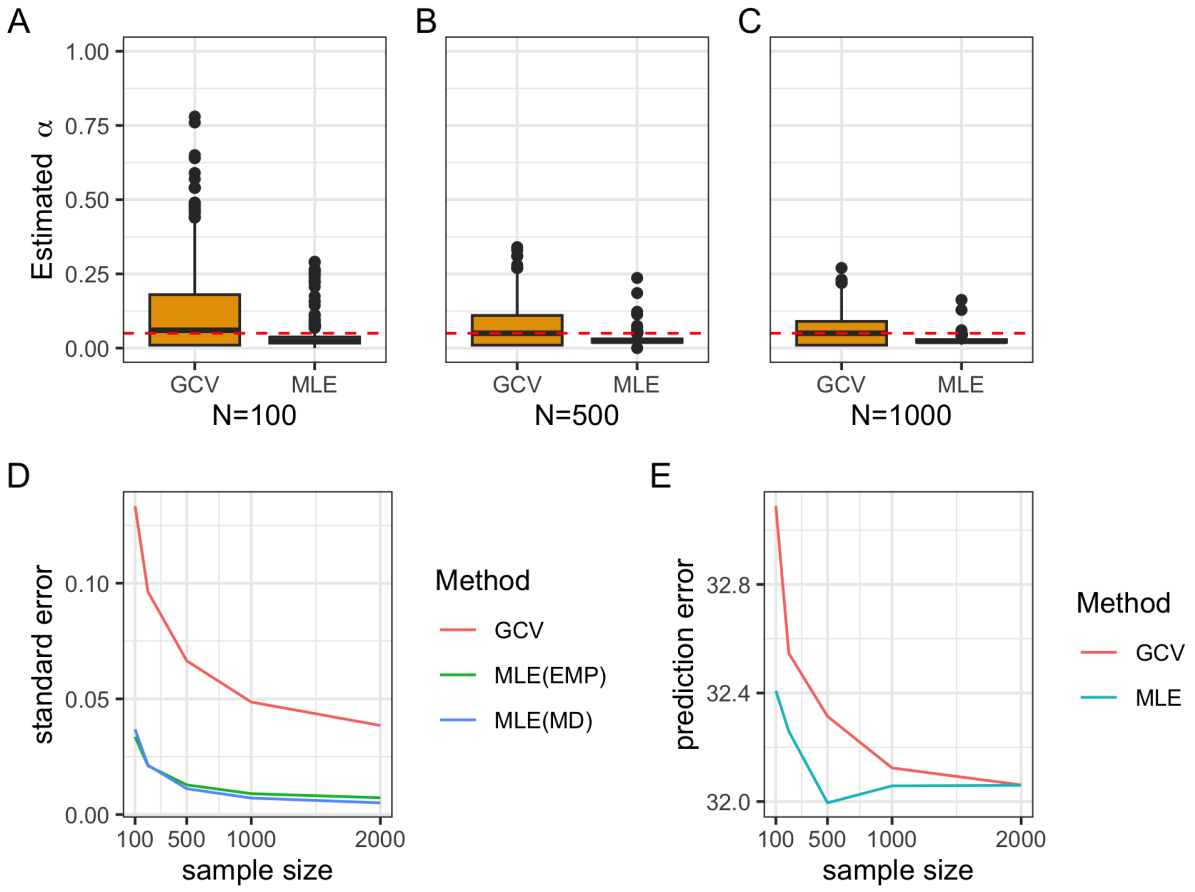

Figure 5.4: Simulation results under **Scenario II**: true $\alpha = 0.05$. **A-C**: Estimation results of $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$ with sample size $N = 100, 500, 1000$. Each box plot showed the results under a unique setting with a certain sample size. N increases from left to right for those plots. **D**: Standard errors of $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$. The empirical-based standard errors of GCV, and both the empirical and model-based standard errors of MLE were considered. **E**: Prediction performances, as measured with root mean squared prediction error, of α-transformed regression model with GCV and MLE. *Abbrevation: Empirical-based, EMP. Model-based, MD.

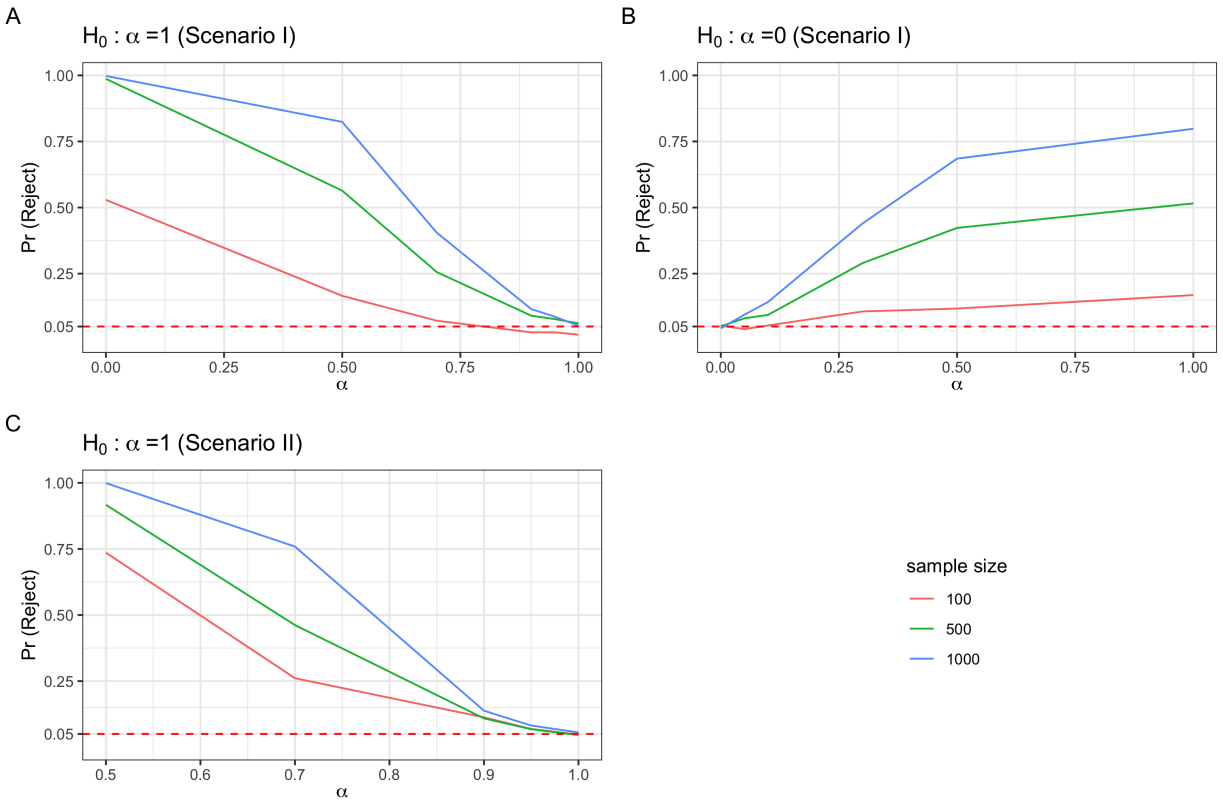

Figure 5.5: Estimated proportions of rejecting the null hypothesis for testing α as indicated in 3.16. **A**: Testing $H_0 : \alpha = 1$, with $\alpha = 0, 0.5, 0.7, 0.9, 0.95, 1$ under alternatives **Scenario I**. **B**: Testing $H_0 : \alpha = 0$, with $\alpha = 0, 0.05, 0.1, 0.3, 0.5, 1$ under alternatives under **Scenario I**. **C**: Testing $H_0 : \alpha = 1$, with $\alpha = 0.5, 0.7, 0.9, 0.95, 1$ under alternatives under **Scenario II**.

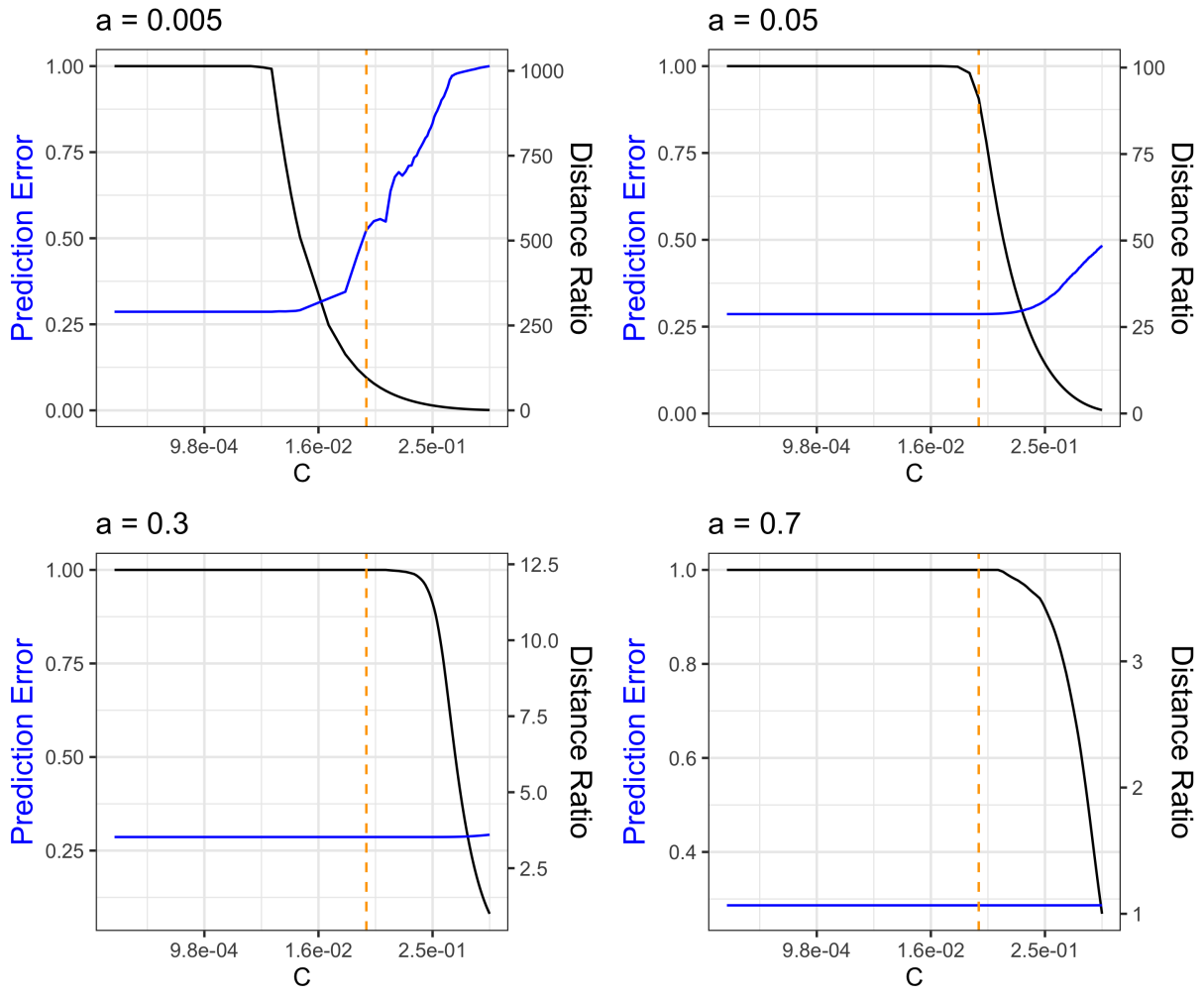

Figure 5.6: Performances of constrained maximum likelihood approach: scaled prediction error $\widetilde{\text{RMSPE}}(c, \alpha_t)$ and the distance ratio metric $R(c)$ with respect to c and true α, α_t, with fixing the sample size $N = 1000$. Each figure corresponds to a specific α_t, which can take on values of 0.005, 0.05, 0.3, and 0.7, while keeping all other settings the same. Within each figure, the left y-axis represents the scaled prediction error $\widetilde{\text{RMSPE}}(c, \alpha_t)$ and the right y-axis represents the distance ratio metric $R(c)$, while the x-axis represents the constraint c, ranging from 1×10^{-4} to 1. The orange dashed line indicates the recommended constraint of $c = 0.05$.

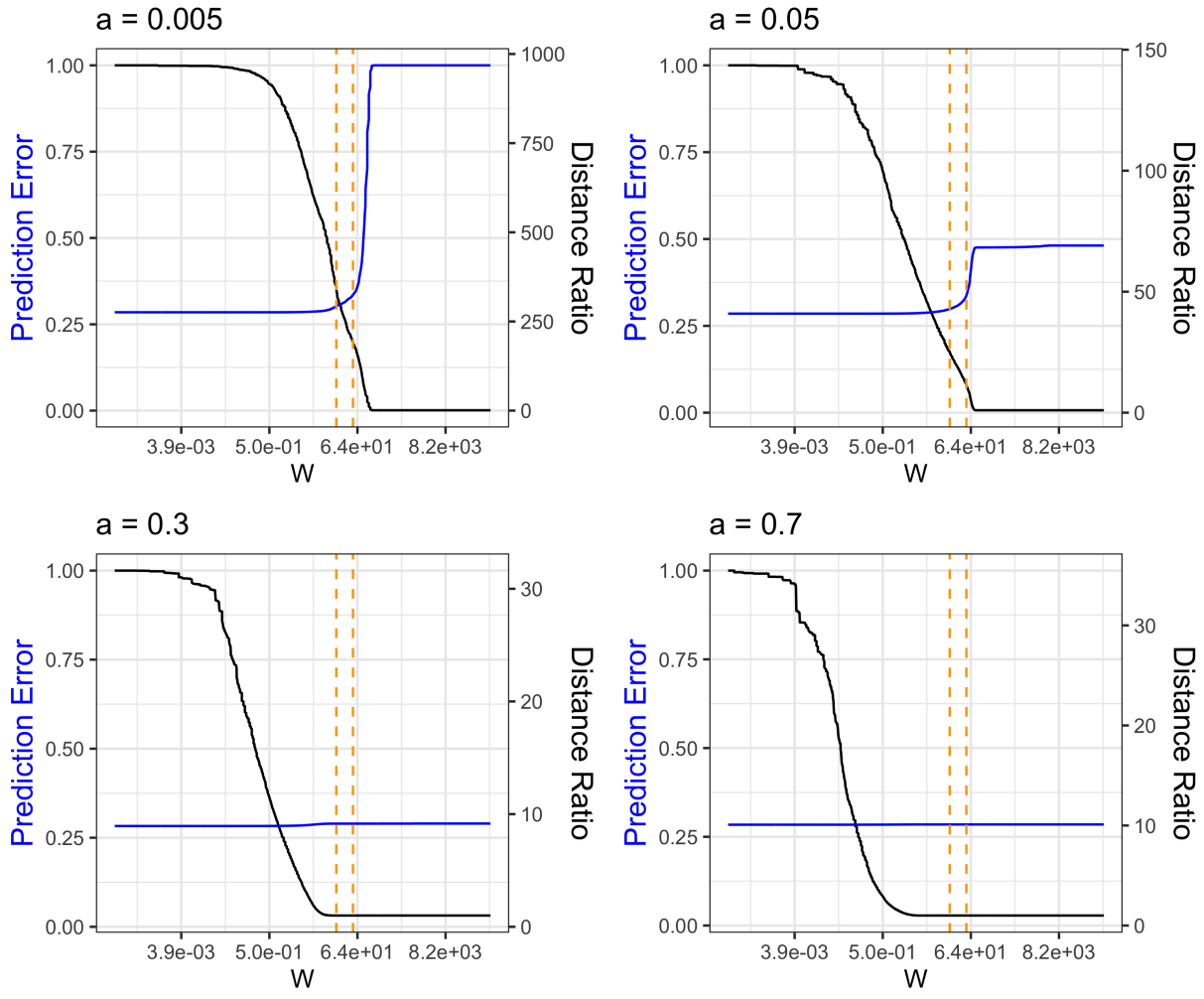

Figure 5.7: Performances of modified likelihood approach: scaled prediction error $\widetilde{\text{RMSPE}}(W, \alpha_t)$ and the distance ratio metric $R(W)$ with respect to W and true α, α_t, with fixing the sample size $N = 100$. Each figure corresponds to a specific α_t, which can take on values of 0.005, 0.05, 0.3, and 0.7, while keeping all other settings the same. Within each figure, the left y-axis represents the scaled prediction error $\widetilde{\text{RMSPE}}(W, \alpha_t)$ and the right y-axis represents the distance ratio metric $R(W)$, while the x-axis represents the regularization coefficient W, ranging from 1×10^{-4} to 1×10^{5}. The orange dashed line indicates the recommended range of $W \in [20, 60]$.

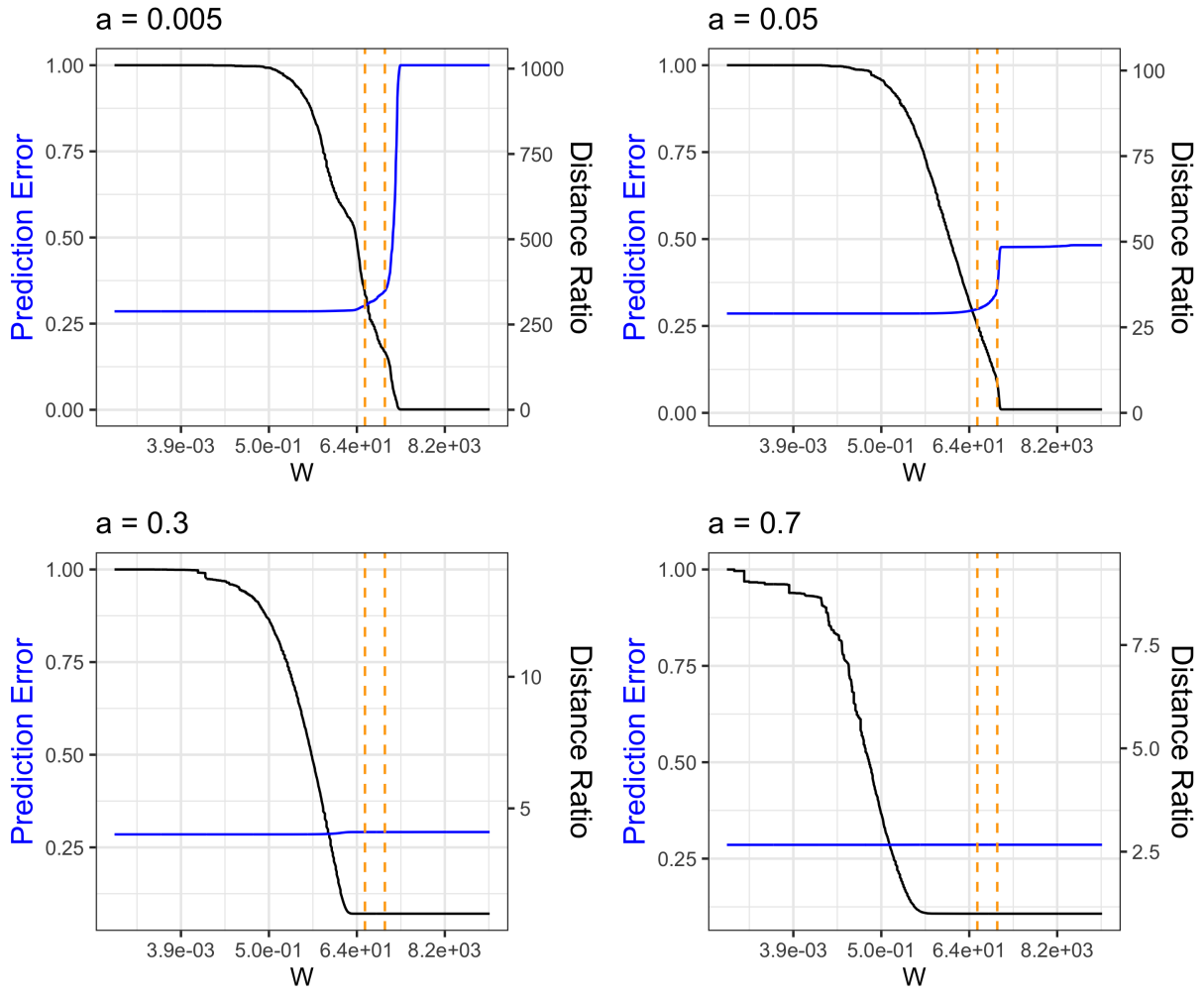

Figure 5.8: Performances of modified likelihood approach: scaled prediction error $\widetilde{\text{RMSPE}}(W, \alpha_t)$ and the distance ratio metric $R(W)$ with respect to W and true α, α_t, with fixing the sample size $N = 500$. Each figure corresponds to a specific α_t, which can take on values of 0.005, 0.05, 0.3, and 0.7, while keeping all other settings the same. Within each figure, the left y-axis represents the scaled prediction error $\widetilde{\text{RMSPE}}(W, \alpha_t)$ and the right y-axis represents the distance ratio metric $R(W)$, while the x-axis represents the regularization coefficient W, ranging from 1×10^{-4} to 1×10^{5}. The orange dashed line indicates the recommended range of $W \in [100, 300]$.

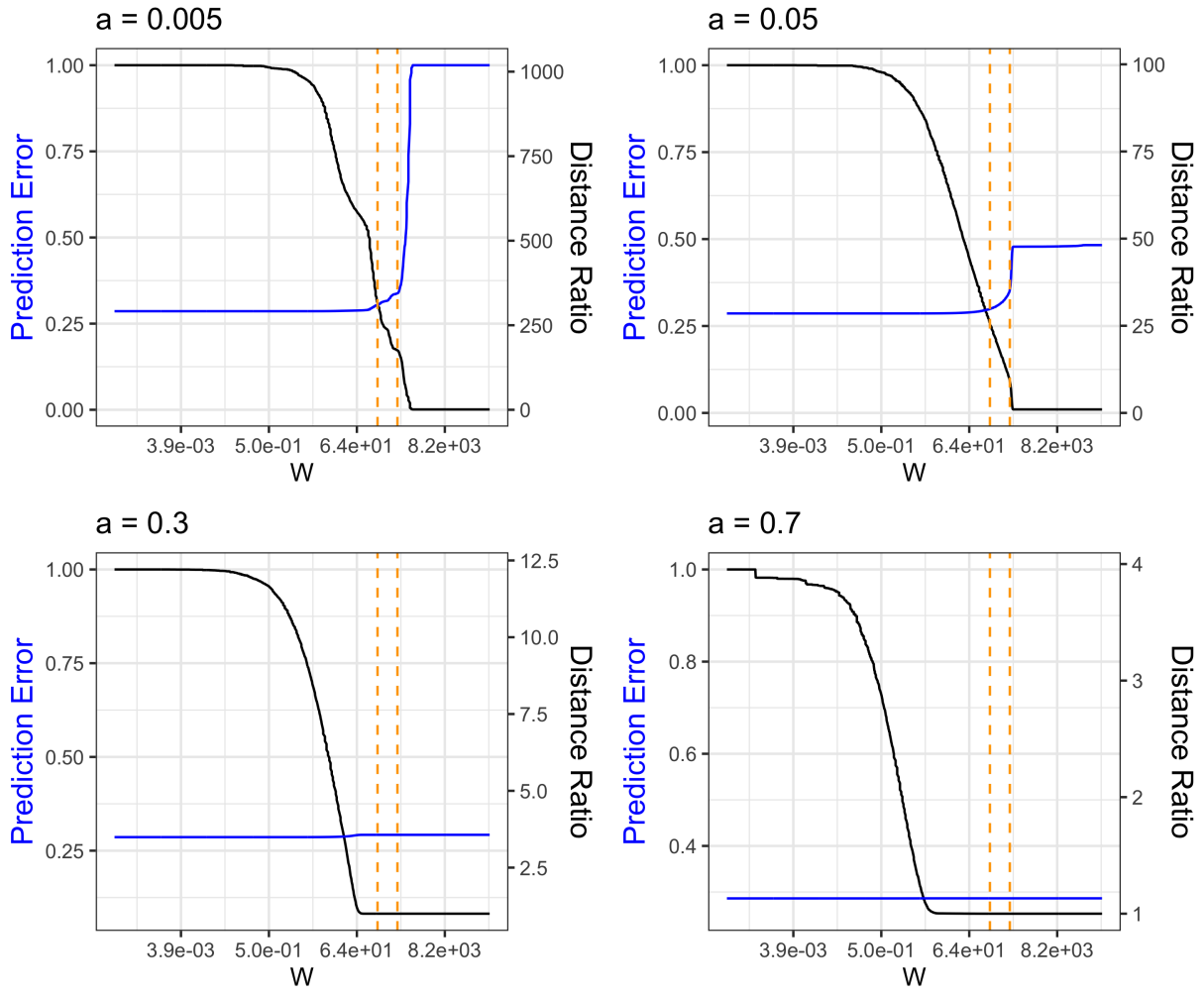

Figure 5.9: Performances of modified likelihood approach: scaled prediction error $\widetilde{\text{RMSPE}}(W, \alpha_t)$ and the distance ratio metric $R(W)$ with respect to W and true α, α_t, with fixing the sample size $N = 1000$. Each figure corresponds to a specific α_t, which can take on values of 0.005, 0.05, 0.3, and 0.7, while keeping all other settings the same. Within each figure, the left y-axis represents the scaled prediction error $\widetilde{\text{RMSPE}}(W, \alpha_t)$ and the right y-axis represents the distance ratio metric $R(W)$, while the x-axis represents the regularization coefficient W, ranging from 1×10^{-4} to 1×10^5. The orange dashed line indicates the recommended range of $W \in [200, 600]$.

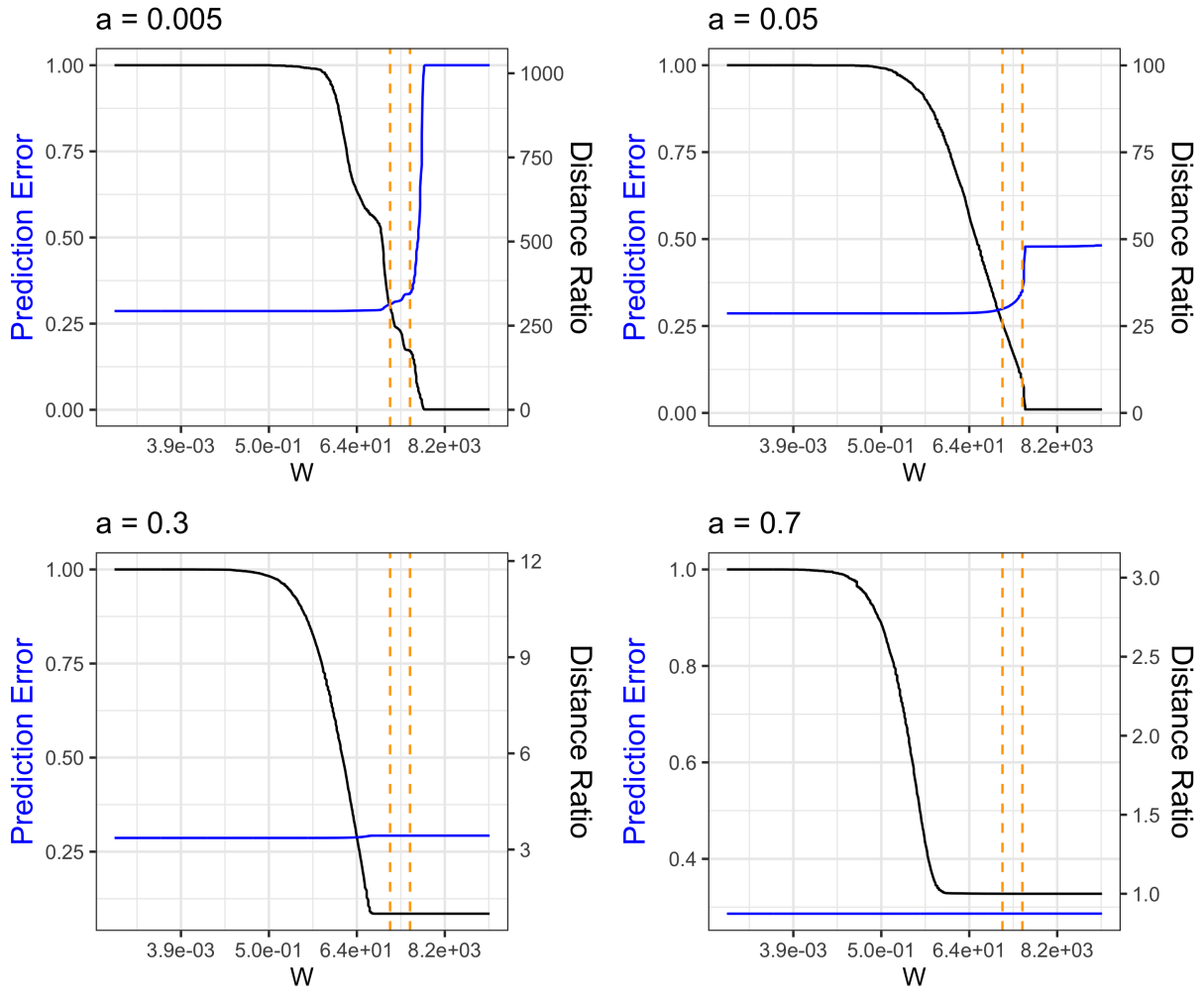

Figure 5.10: Performances of modified likelihood approach: scaled prediction error $\widetilde{\text{RMSPE}}(W, \alpha_t)$ and the distance ratio metric $R(W)$ with respect to W and true α, α_t, with fixing the sample size $N = 2000$. Each figure corresponds to a specific α_t, which can take on values of 0.005, 0.05, 0.3, and 0.7, while keeping all other settings the same. Within each figure, the left y-axis represents the scaled prediction error $\widetilde{\text{RMSPE}}(W, \alpha_t)$ and the right y-axis represents the distance ratio metric $R(W)$, while the x-axis represents the regularization coefficient W, ranging from 1×10^{-4} to 1×10^{5}. The orange dashed line indicates the recommended range of $W \in [400, 1200]$.

Chapter 6

APPLICATION TO NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

In this section, we embarked upon a comprehensive examination of the NHANES dataset. The NHANES dataset is a comprehensive collection through various demographic, dietary, examination, laboratory, questionnaire, and other measurements to assess the health and nutritious status of a representative sample of the United States population. The ethics committee of the Centers for Disease Control and Prevention (CDC) approved the study, and the detailed study procedures and methods are available on the CDC website. Physical activity measurements were conducted and were publicly available from the NHANES data of the 2005-2006 subset. We begin with an exploratory analysis. Subsequently, our analysis involved the implementation of multiple methodologies, including ISM, ILR-transformed regression, and power transformation-based regression.

6.1 Exploratory Analysis

The NHANES dataset (National Center for Health Statistics [2005]) comprised a total of N=1,333 observations. In our study, we specifically focused on glucose measurement (mg/dL) as the outcome biomarker, and examined the association between glucose and different physical activity behaviors, including SB, LPA and MVPA, as covariates of interest. We adjusted for several factors during the analysis, including age (years), gender, total saturated fatty acids (gm), daily caffeine intake (mg), and daily energy intake (kcal). Figure 6.2 displayed both histograms and scatter plots depicting the distribution of SB, LPA, MVPA, and glucose, as well as their associations. Notably, the majority of participants demonstrated glucose levels being around 5 mg/dL. The data also indicated that, for most participants, sedentary behavior constituted a significant portion of their daily activities, spanning a majority of the 24-hour timeframe. Light physical activity also accounted for a substantial portion of their time, although slightly less than sedentary behavior. On the other

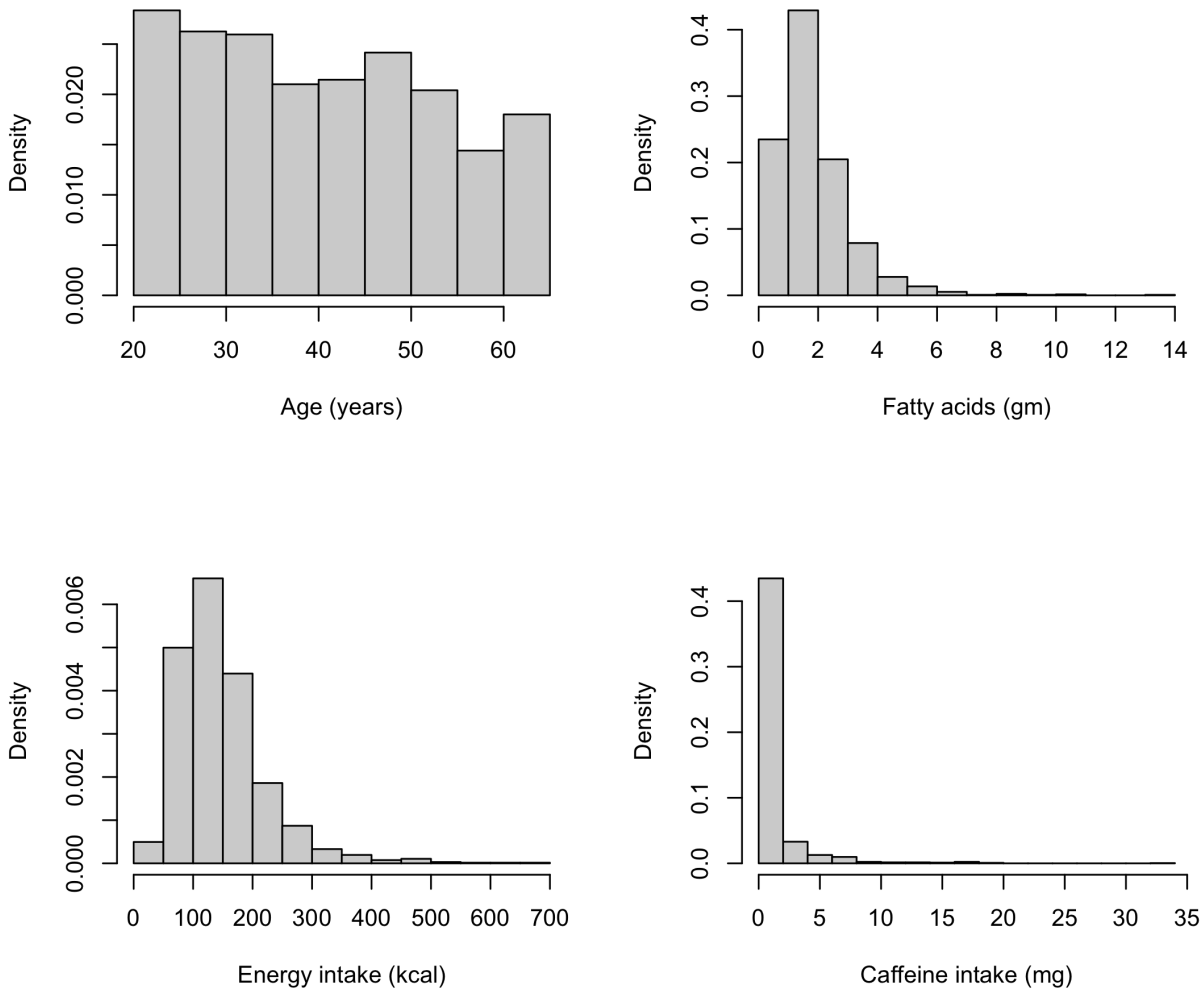

Figure 6.1: Histograms of covaraites for NHANES (N=1333).

hand, moderate-to-vigorous physical activity only occupied a small fraction of their daily routine. The strong association between SB and LPA can be attributed to the inherent constraint of the 24-hour timeframe, which is also supported by their correlation presented in Figure 6.2. It was worth noting that a small proportion of participants in the NHANES dataset did not have any recorded daily activity in the categories of LPA (N=3) or MVPA (N=19), and subjects who have 0 minutes spent in the LPA category also do not have any time spent in the MVPA category. Within the

Table 6.1: Summary statistics of variables for NHANES (N=1333).

	Min	Q1	Mean	Median	Q3	Max	SD
Age (years)	21	30	40	40.876	51	64	12.768
Gender	-	-	0.53	-	-	-	-
Fatty acids (gm)	0.073	1.036	1.545	1.846	2.347	13.440	1.249
Caffeine intake (mg)	0	0	0	0.837	0.022	33.475	2.439
Energy intake (kcal)	6.750	96.629	133.859	147.624	180.394	664.800	75.524
SB	0.277	0.627	0.712	0.717	0.808	1	0.135
LPA	0	0.178	0.271	0.264	0.347	0.712	0.126
MVPA	0	0.005	0.014	0.020	0.027	0.166	0.020
$Z_1^{(\hat{\alpha}_c)}$	-0.866	0.527	0.527	1.090	1.375	6.428	0.966
$Z_2^{(\hat{\alpha}_c)}$	0.710	1.864	2.182	2.205	2.528	3.711	0.490
Glucose (mg/dL)	2.498	4.940	5.329	5.676	5.773	23.203	1.852

dataset, there were 627 male participants (coded as 1) and 706 female participants (coded as 2), and we re-coded males as 0 and females as 1 in the following analysis. Figure 6.1 presented a histogram illustrating the distribution of the adjusted continuous covariates, which include age, fatty acids, caffeine intake, and energy intake. The histogram revealed a relatively similar distribution of individuals across different age categories, although there was a higher proportion of younger individuals compared to older ones. Additionally, the majority of participants exhibited fatty acid levels ranging from 0 to 3 gm, energy intake ranging from 50 to 200 kcal, and caffeine intake ranging from 0 to 3 mg. More summary statistics of these variables were shown in Table 6.1.

6.2 Regression Analysis

Chastin et al. [2015] previously performed association analysis with both isotemporal and ILR-transformed regression approaches and concluded that ILR-transformed regression indicates unequal PA category substitution effect on several biomarkers treated as the model outcomes. The

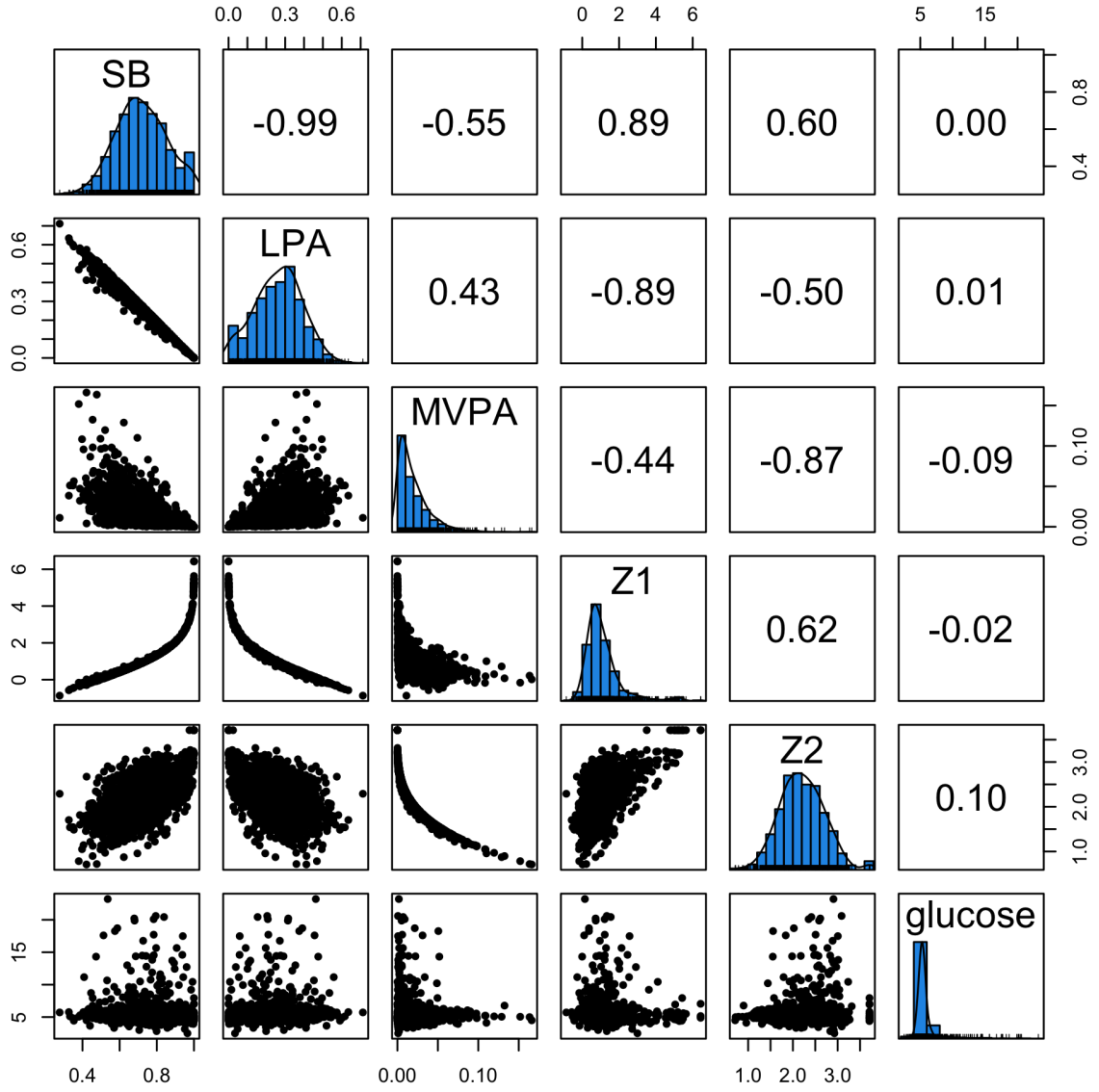

Figure 6.2: Histogram and correlations of variables of interest for NHNES (N=1333). Z_1 and Z_2 represent for $Z_1^{(\hat{\alpha}_c)}$ and $Z_2^{(\hat{\alpha}_c)}$, respectively, where $\hat{\alpha}_c = 0.33$.

NAHANES data PA levels measured by accelerometers were classified into three PA categories (SB, LPA, and MVPA) using activity count per minute thresholds, averaged over all available days, and re-scaled to the proportions of 24 hours a day spent in each category. The PA categories were

treated as compositional predictors in the regression analyses, and adjusted with several other measured factors to study their association with health outcomes such as BMI, waist circumference, and other biomarkers.

In our study, we applied the ISM, ILR-transformed as well as PTR to the NAHANES data collected during 2005-2006. The ISM with the following expression (6.1) estimated the effect of replacing activities in one intensity with activities in another intensity for the same amount of time. Specifically, we included SB and LPA as the primary behaviors.

$$\begin{aligned}\text{Glucose} = \beta_0^{ISM} + \beta_1^{ISM}\text{SB} + \beta_2^{ISM}\text{LPA} + \beta_3^{ISM}\text{Age} \\ + \beta_4^{ISM}\text{Gender} + \beta_5^{ISM}\text{Acid} + \beta_6^{ISM}\text{Caffeine} + \beta_7^{ISM}\text{Energy}.\end{aligned} \tag{6.1}$$

When conducting ILR-transformed regression analysis, there were two possible approaches for dealing with subjects who had zero compositions: either disregarding these subjects altogether or substituting their zero compositions with a small positive value. Although both methods raised concerns as they can significantly affect the ILR-transformed data, we opted to replace the zero physical activity values with a small value, specifically 0.001 minutes. This choice was made to preserve the sample size across all methods. The ILR-transformed regression model can be expressed as follows:

$$\begin{aligned}\text{Glucose} = \beta_0^{ILR} + \beta_1^{ILR}\text{Z}_1^{ILR} + \beta_2^{ILR}\text{Z}_2^{ILR} + \beta_3^{ILR}\text{Age} \\ + \beta_4^{ILR}\text{Gender} + \beta_5^{ILR}\text{Acid} + \beta_6^{ILR}\text{Caffeine} + \beta_7^{ILR}\text{Energy},\end{aligned} \tag{6.2}$$

where Z_1^{ILR} and Z_2^{ILR} are the first and second column of the ILR-transformed data Z^{ILR}, respectively.

By contrast, PTR did not require non-zero values for physical activity components. Therefore, it can be directly applied to NHANES data without any special treatment for subjects with zero physical activity values. The PTR model was shown in (6.3).

$$\begin{aligned}\text{Glucose} = \beta_0^{PTR} + \beta_1^{PTR}\text{Z}_1^{(\alpha)} + \beta_2^{PTR}\text{Z}_2^{(\alpha)} + \beta_3^{PTR}\text{Age} \\ + \beta_4^{PTR}\text{Gender} + \beta_5^{PTR}\text{Acid} + \beta_6^{PTR}\text{Caffeine} + \beta_7^{PTR}\text{Energy},\end{aligned} \tag{6.3}$$

where $\text{Z}_1^{(\alpha)}$ and $\text{Z}_2^{(\alpha)}$ were the first and second column of the power transformed data $\text{Z}^{(\alpha)}$, respectively. In our analysis, we utilized both the constrained maximum likelihood and modified likelihood approach for PTR, and set the constraint $c = 0.05$ and the penalty coefficient $W = 1$.

In the real data analysis, the selection of W was approached using a parametric bootstrap-based method. We employed the estimated results from the constrained maximum likelihood approach for PTR, as presented in Table 6.2, to set the true values of α and the coefficients β. To ensure consistency, we denoted the true standard deviation of glucose as σ_N and considered a sample size of $N = 1333$, which matched the size of the real dataset. Following the methodology outlined in Section 5.4, we computed the prediction error, RMSEP(W), and the distance ratio, $R(W)$, for different regularization coefficients W using the modified likelihood approach. Examining the outcomes depicted in Figure 6.3, we observed that the prediction error exhibited minimal sensitivity to changes in W. Consequently, selecting $W = 1$ appeared reasonable since, at that juncture, the distance ratio had already decreased significantly while the prediction error remained relatively unchanged.

6.3 Results

The results of model fitting for the three models were presented in Table 6.2. Regarding the implementation of the proposed constrained maximum likelihood approach for the PTR model, the estimated value of α is $\hat{\alpha}_c = 0.33$. In the case of the modified likelihood approach, the estimated α was $\hat{\alpha}_m = 0.39$, which was in close proximity to the estimate obtained through the constrained maximum likelihood approach. Given the similarity in the estimates of α obtained by these two estimation procedures, the estimated coefficients also exhibited resemblance, as presented in Table 6.2. The prediction performances, assessed using the cross-validated sum of squared predicted error (SSPE), were presented in Table 6.3. The table clearly indicated that the ILR-transformed method demonstrated the highest SSPE, followed by the ISM approach with a slightly smaller error. By contrast, the proposed PTR model exhibited the lowest error, thereby underscoring its superiority over the other two methods in terms of prediction performance. This was not surprising, as the family of PTR models was more general and included both ILR and ISM as special cases. The two estimation methods, the constrained estimation and the modified likelihood approach, yield comparable prediction performances.

We also explored the association between the transformed data and the original PA behaviors data (Figure 6.2). The figure revealed a non-linear relationship between SB and LPA with $Z_1^{(\hat{\alpha}_c)}$, as

Table 6.2: Model fitting results of ISM, ILR-transformed and PTR as indicated in (6.1), (6.2), (6.3). *Abbreviation: Constrained maximum likelihood approach for PTR, PTR-C. modified likelihood approach for PTR, PTR-M.

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$
ISM	Coefficient	8.621	8.530	0.041	-0.357	0.054	-0.026	-0.001
	Std. Error	2.842	3.072	0.004	0.112	0.071	0.024	0.001
	p-value	0.002	0.006	< 0.001	0.001	0.449	0.285	0.342
ILR	Coefficient	-0.031	0.093	0.041	-0.297	0.066	-0.024	-0.001
	Std. Error	0.058	0.040	0.004	0.109	0.071	0.024	0.001
	p-value	0.593	0.02	< 0.001	0.006	0.351	0.326	0.286
PTR-C	Coefficient	-0.110	0.577	0.038	-0.413	0.060	-0.021	-0.001
	Std. Error	0.069	0.140	0.004	0.113	0.071	0.024	0.001
	p-value	0.112	< 0.001	< 0.001	< 0.001	0.396	0.394	0.260
PTR-M	Coefficient	-0.100	0.671	0.038	-0.413	0.059	-0.021	-0.001
	Std. Error	0.072	0.165	0.004	0.113	0.071	0.024	0.001
	p-value	0.165	< 0.001	< 0.001	< 0.001	0.408	0.382	0.270

Table 6.3: Cross-validated sum of squared prediction error (SSPE) of ISM, ILR-transformed and PTR as indicated in (6.1), (6.2), (6.3). *Abbreviation: Constrained maximum likelihood approach for PTR, PTR-C. modified likelihood approach for PTR, PTR-M.

	ISM	ILR	PTR-C	PTR-M
MSPE	3915.30	3936.03	3885.71	3886.03

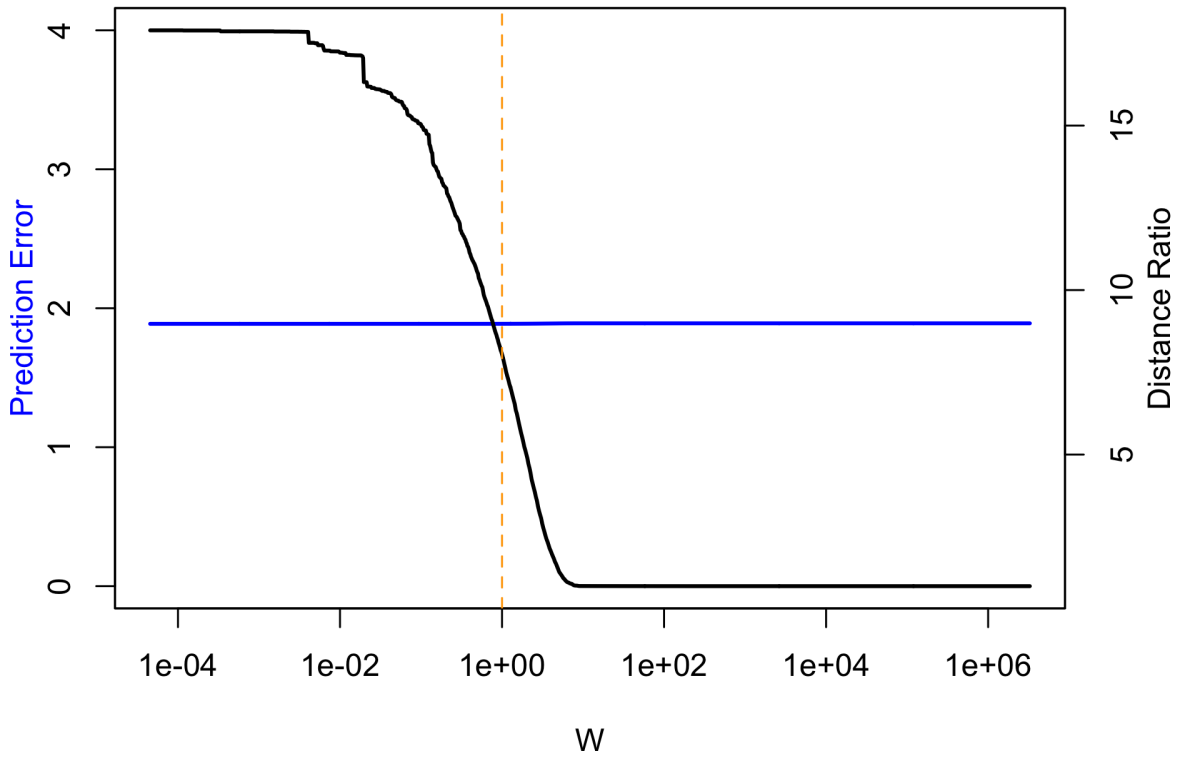

Figure 6.3: Performances of modified likelihood approach: prediction error RMSPE(W) and the distance ratio metric $R(W)$ with respect to W. The left y-axis represents the prediction error and the right y-axis represents the distance ratio metric $R(W)$, while the x-axis represents the regularization coefficient W, ranging from 1×10^{-4} to 1×10^{5}. The orange dashed line indicates the recommended W=1.

well as a non-linear relationship between MVPA and $Z_2^{(\hat{\alpha}_c)}$. This observation aligns with the formula of the power transformation described in (3.2) and (2.2). Furthermore, the inclusion of highly correlated covariates in the regression model can lead to collinearity issues. However, by examining the figure, it became evident that the application of power transformation effectively mitigated the previously observed high correlation between SB and LPA. This observation was supported by

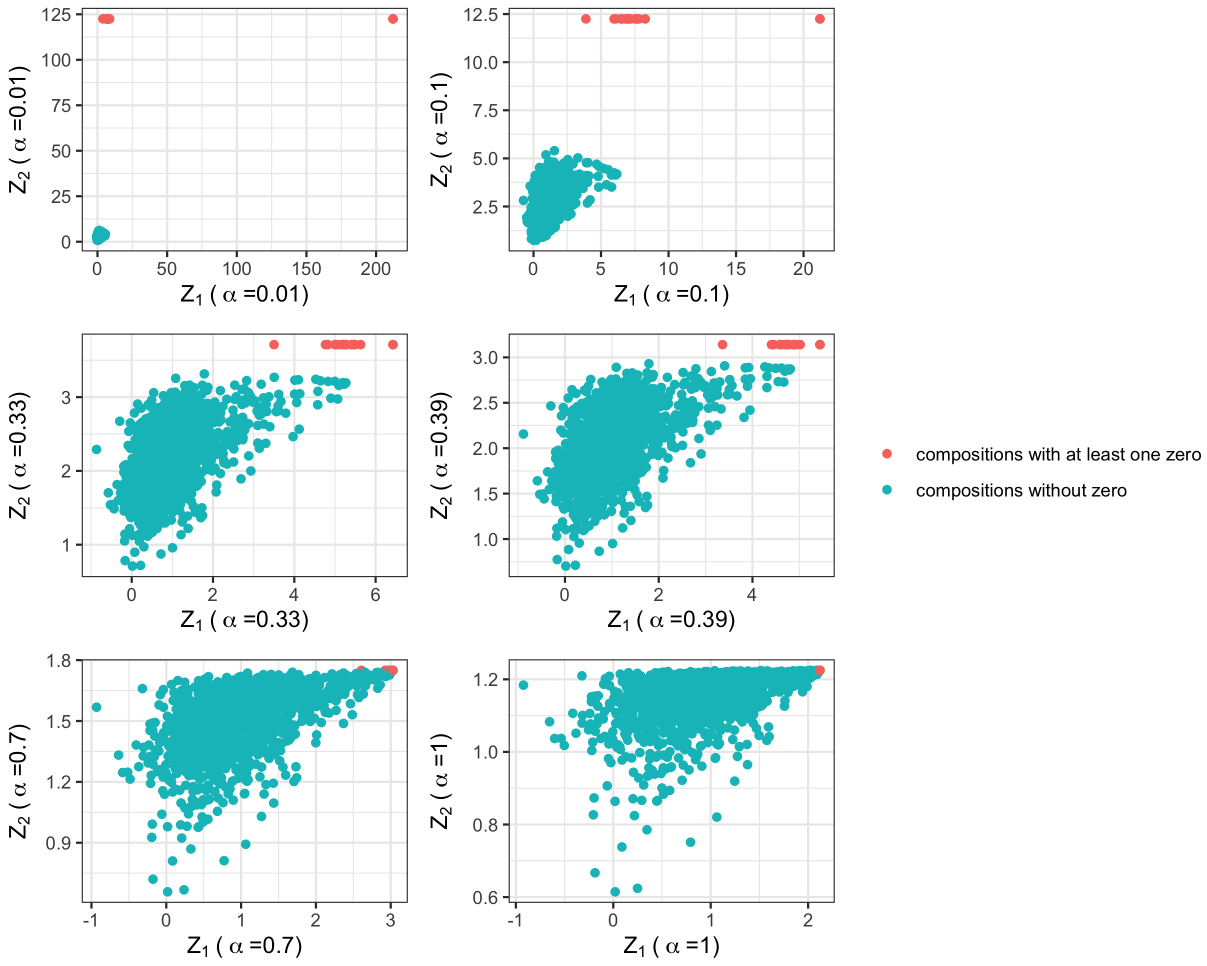

Figure 6.4: The first column of the transformed data, $Z_1^{(\alpha)}$, versus the second column of the transformed data, $Z_2^{(\alpha)}$, considering different α, including $\alpha = 0.01, 0.1, \hat{\alpha}_c(=0.33), \hat{\alpha}_m(=0.39)$, 0.7, 1.

the correlation between $Z_1^{(\hat{\alpha}_c)}$ and $Z_2^{(\hat{\alpha}_c)}$. Regarding the power transformation, we also investigated the impact of different values of α on the transformed data (Figure 6.4). It can be observed that for observations without zero values, varying α does not significantly affect the transformation of these data. However, for observations with zero compositions, smaller values of α, such as 0.01 or 0.1, resulted in very large transformed values, effectively creating outliers when compared to the transformed data for non-zero compositional data. Based on Figure 6.4, it was worth noting that

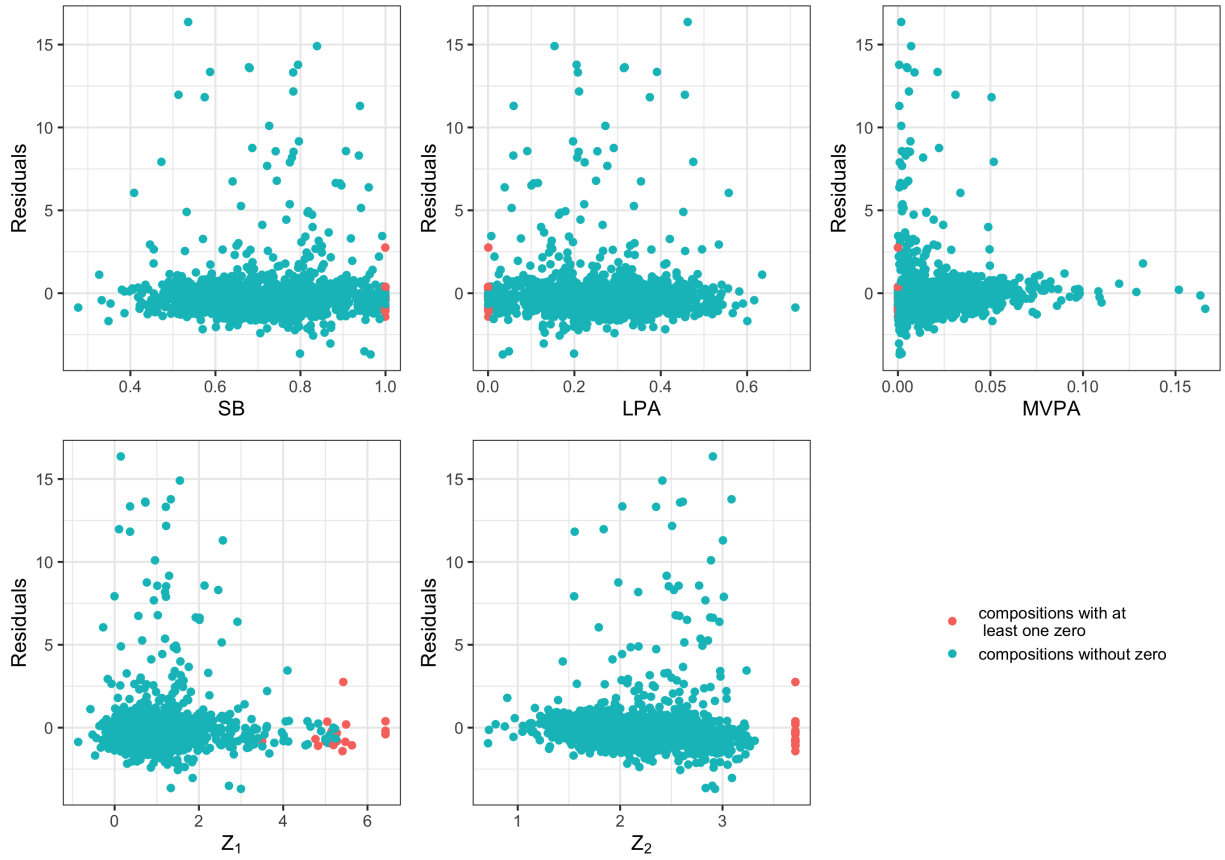

Figure 6.5: Residuals versus different covariates for PTR with constrained maximum likelihood approach. Z_1 and Z_2 represent for $Z_1^{(\hat{\alpha}_c)}$ and $Z_2^{(\hat{\alpha}_c)}$, respectively.

the application of power transformation with $\hat{\alpha}_c$ and $\hat{\alpha}_m$ yielded comparable results between transformed observations with zeros and those without zeros. This finding suggested that the presence of zero values was unlikely to exert a significant impact on the fitting of the model.

In the subsequent discussion, we primarily focused on the results obtained from the constrained maximum likelihood approach due to its similar performance to the modified likelihood approach in PTR Figure 6.5 illustrated the relationship between the residuals and the covariates of interest, including SB, LPA and MVPA, as well as the power transformed data, $Z_1^{(\hat{\alpha}_c)}$ and $Z_2^{(\hat{\alpha}_c)}$. It was evident that the residuals exhibited a random distribution around zero. Notably, observations with zero compositions displayed a similar pattern, suggesting that these particular observations do not

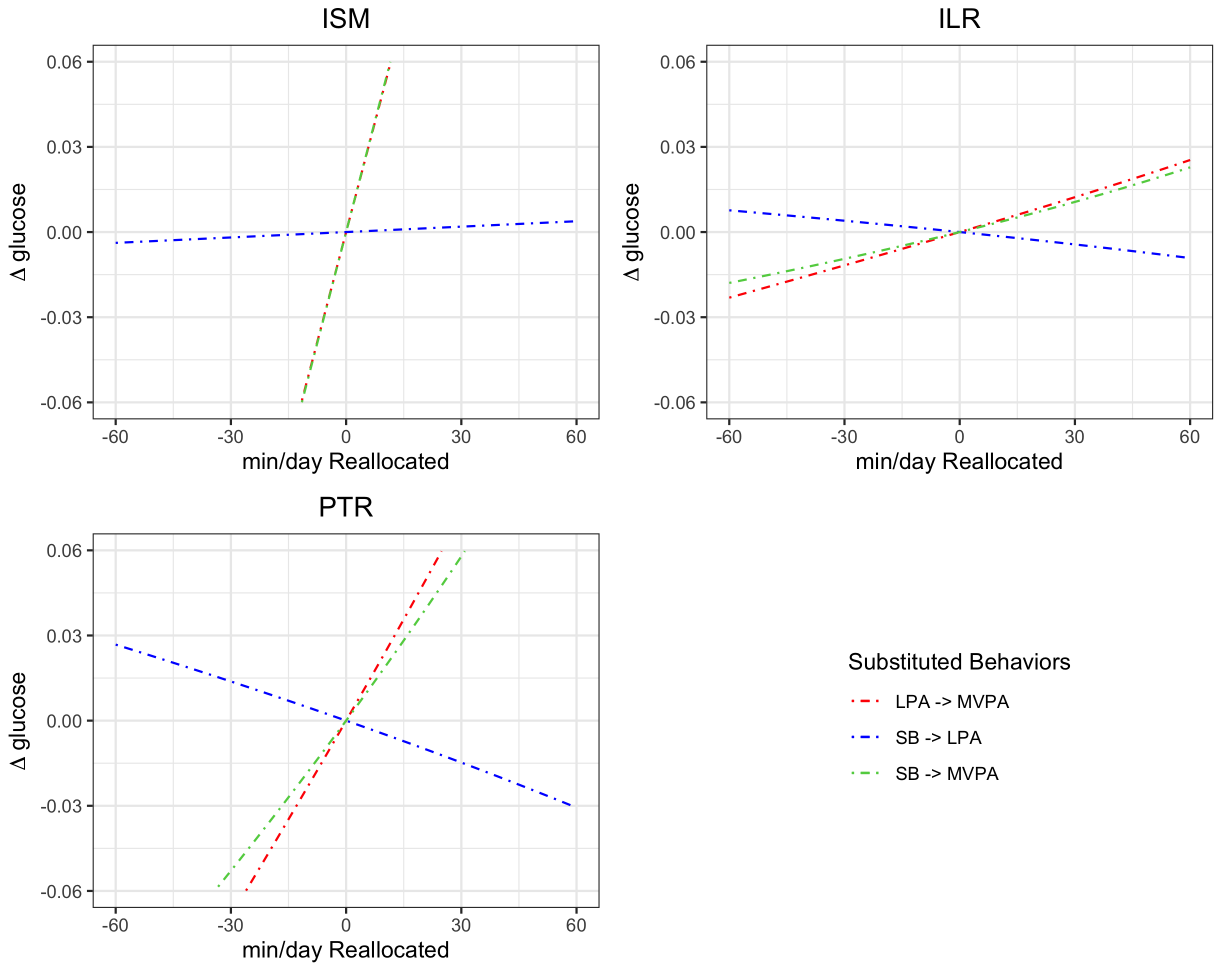

Figure 6.6: Estimated substitution effect with ISM (left upper), ILR-transformed regression (right upper), and PTR with constrained maximum likelihood approach (left bottom).

pose any issues in terms of model fitting. This outcome aligned with our expectations since the estimated value of α, $\hat{\alpha}_c = 0.33$, is relatively large.

The substitution effect of various physical activity behaviors using the three methods was illustrated in Figure 6.6. Regarding ISM, the substitution effects between LPA and MVPA, as well as SB and MVPA, were quite similar. However, they differed significantly from the effect observed between LPA and SB. When applying the ILR-transformed regression, the effects of substituting LPA with MVPA and SB with MVPA demonstrated similarity. However, the disparities became

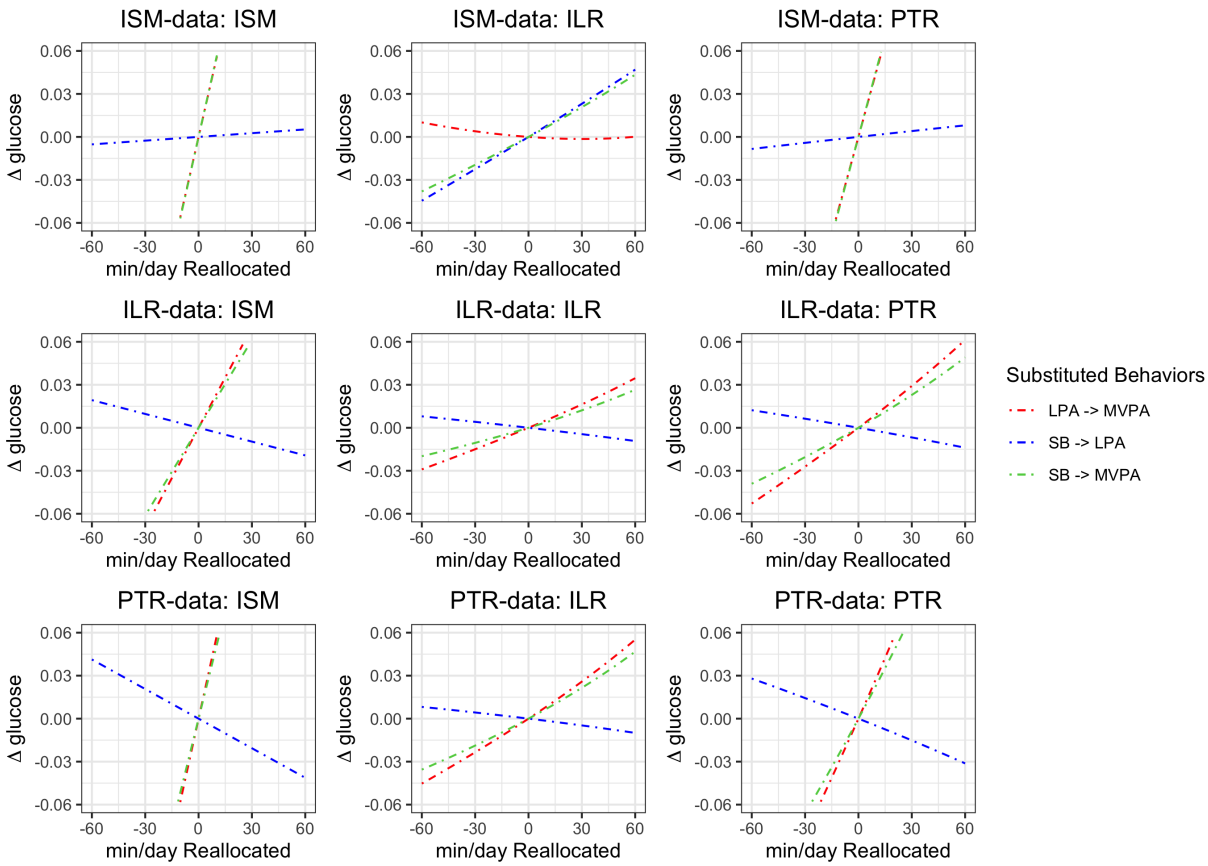

Figure 6.7: Estimated substitution effect with ISM, ILR-transformed regression, and PTR under different data generation mechanisms.

more pronounced when examining the effects in the opposite direction of these behaviors. Additionally, the effects between SB and LPA deviated from those observed between LPA and MVPA, as well as SB and MVPA. The employment of PTR revealed close substitution effects between LPA and MVPA, as well as SB and MVPA. Nevertheless, these effects diverged significantly from the observed effect between LPA and SB.

To investigate the substitution effects using different methods and ensure the comparability of results, we employed a parametric bootstrap-based method. Three data generation mechanisms were considered based on the fitted results presented in Table 6.2. By considering the estimated results of ISM, ILR, and PRT-C in Table 6.2 as the true values, we generated data using these three

methods. Subsequently, we applied each method to the generated data under each mechanism to investigate the substitution effects, as depicted in Figure 6.7. The figure illustrated that each row corresponds to a specific data generation mechanism, namely ISM, ILR, and PTR-data. To successfully implement ILR, we replaced zeros with small values (0.01/1440) when employing this method. Additionally, the constrained maximum likelihood approach was used for PTR to handle data with zero compositions. Under the ISM-based data generation mechanism, we observed that PTR yielded nearly identical results to ISM, which represented the true substitution effect in this scenario. Conversely, ILR produced significantly different outcomes. This outcome aligned with expectations, as ISM can be considered a special case of PTR. Considering the data generated based on ILR, we first replaced zeros with 0.01/1440 and then applied the three methods. Since no zero compositions were present in this case, we used the MLE method proposed in Section 3.2 for PTR. Despite PTR and ISM not accurately estimating the true effect, it was evident that PTR outperformed ISM and exhibited greater similarity to the true effects. Regarding data generated based on PTR, although ISM outperformed ILR in estimating the substitution effects by being closer to the true effects, neither method accurately estimated the true effects. In conclusion, the proposed PTR method surpassed ILR and ISM in estimating the substitution effects between different PA behaviors.

Chapter 7

DISUCSSION

7.1 Summary

In this thesis, we proposed a novel power transformation-based compositional data analysis approach and conducted extensive investigations. To begin with, we examined the orthogonality property of the power transformation and presented rigorous proof. Subsequently, we first considered the scenario in the absence of zero values in compositional data and discussed two estimation approaches, maximum likelihood and generalized cross-validation. Furthermore, we presented asymptotic properties of the ML-based estimator of α. We explored scenarios where the value of α resides on the boundary of its parameter space, which violated standard regularity conditions. Moreover, we developed a bootstrap-based confidence interval for predicting outcomes, which accounts for uncertainty in estimating the tuning parameter and exhibits asymptotic invariance to the distribution of observations.

Next, we extended the proposed method for modeling compositional data that includes zero values. In this scenario, the power transformation can result in extreme values or outliers in transformed variables, especially when α is small. To address this issue, we presented two strategies. The initial strategy involves employing constrained maximum likelihood approach, wherein a positive constraint c was introduced to confine α within the range of $[c, 1]$. The second method entails maximizing a modified likelihood function, which includes a penalty on α to discourage it from being small. Both of these strategies involved a trade-off between prediction performance and the impact of α on the data transformation.

In the simulation studies, we initially investigated the finite sample properties and model-fitting performances of both the MLE and the GCV-based estimators. Based on the simulation results, our recommendation favors the MLE over the GCV due to its consistent performance across various simulation scenarios. Furthermore, we conducted comprehensive studies to evaluate the perfor-

mance of the proposed constrained maximum likelihood approach and modified likelihood approach. Additionally, we provided specific recommendations regarding the selection of tuning parameters for these two methods. To further validate our findings, we implemented two commonly utilized approaches, the ISM and the ILR-transformed regression. We compared their prediction performance with that of the two proposed methods, demonstrating the superior performance of the proposed methods, particularly when the true value of α is relatively small. Finally, we employed diverse methods to analyze the NHANES dataset, ensuring a thorough examination of the data.

7.2 Future Works

There exist several potential limitations and opportunities for future research in our study. Firstly, regarding the modified likelihood approach, we treated the penalty coefficient W as a constant, and the choice of W has been solely determined based on its performance in finite samples. Alternatively, one might consider estimating W adaptively from the data. Another promising direction for exploration lies in the development of fully nonparametric regression methods with or without utilizing the power transformation. While the proposed power transformation based model allows nonlinear relationships, the functional form is limited to a parametric nonlinear family. This assumption might be restrictive in some settings. Thus, adopting fully nonparametric models will allow more flexible relatinships between PA compositions and health outcomes.

BIBLIOGRAPHY

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

John Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983.

John Aitchison. The statistical analysis of compositional data. caldwell, 2003.

Colin GG Aitken, Grzegorz Zadora, and David Lucy. A two-level model for evidence evaluation. *Journal of forensic sciences*, 52(2):412–419, 2007.

Abdulaziz Alenazi. A review of compositional data analysis and recent advances. *Communications in Statistics-Theory and Methods*, pages 1–33, 2021.

Divya Ankam and Nizar Bouguila. Compositional data analysis with pls-da and security applications. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 338–345. IEEE, 2018.

J Atchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

Jiawei Bai, Chongzhi Di, Luo Xiao, Kelly R Evenson, Andrea Z LaCroix, Ciprian M Crainiceanu, and David M Buchner. An activity index for raw accelerometry data and its comparison with other activity metrics. *PloS one*, 11(8):e0160644, 2016a.

Yang Bai, Gregory J Welk, Yoon Ho Nam, Joey A Lee, Jung-Min Lee, Youngwon Kim, Nathan F Meier, and Philip M Dixon. Comparison of consumer and research monitors under semistructured settings. *Medicine & Science in Sports & Exercise*, 48(1):151–158, 2016b.

Pooja Bansil, Elena V Kuklina, Robert K Merritt, and Paula W Yoon. Associations between sleep disorders, sleep duration, quality of sleep, and hypertension: results from the national health and nutrition examination survey, 2005 to 2008. *The Journal of Clinical Hypertension*, 13(10): 739–743, 2011.

Carles Barceló-Vidal, José A Martín-Fernández, and Vera Pawlowsky-Glahn. Mathematical foundations of compositional data analysis. In *Proceedings of IAMG*, volume 1, pages 1–20, 2001.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*, 2021.

Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester Mackey. Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, 33:16339–16350, 2020.

John Bear and Dean Billheimer. A logistic normal mixture model for compositional data allowing essential zeros. *Austrian Journal of Statistics*, 45(4):3–23, 2016.

Antonella Buccianti, Barbara Nisi, Josep Antoni Martín-Fernández, and Javier Palarea-Albaladejo. Methods to investigate the geochemistry of groundwaters with values for nitrogen compounds below the detection limit. *Journal of Geochemical Exploration*, 141:78–88, 2014.

Matthew P Buman, Elisabeth AH Winkler, Jonathan M Kurka, Eric B Hekler, Carol M Baldwin, Neville Owen, Barbara E Ainsworth, Genevieve N Healy, and Paul A Gardiner. Reallocating time to sleep, sedentary behaviors, or active behaviors: associations with cardiovascular disease risk biomarkers, nhanes 2005–2006. *American journal of epidemiology*, 179(3):323–334, 2014.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *Siam Journal on Scientific Computing*, 16(5):1190–1208, 1995.

CARL J Caspersen. Physical activity epidemiology: concepts, methods, and applications to exercise science. *Exercise and sport sciences reviews*, 17:423–473, 1989.

Carl J Caspersen, Kenneth E Powell, and Gregory M Christenson. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public health reports*, 100(2):126, 1985.

Jean-Philippe Chaput, Casey E Gray, Veronica J Poitras, Valerie Carson, Reut Gruber, Timothy Olds, Shelly K Weiss, Sarah Connor Gorber, Michelle E Kho, Margaret Sampson, et al. Systematic review of the relationships between sleep duration and health indicators in school-aged children and youth. *Applied physiology, nutrition, and metabolism*, 41(6):S266–S282, 2016.

Sebastien FM Chastin, Javier Palarea-Albaladejo, Manon L Dontje, and Dawn A Skelton. Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. *PloS one*, 10 (10):e0139984, 2015.

Hanfeng Chen, Jiahua Chen, and John D Kalbfleisch. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):19–29, 2001.

Jihyun E Choi and Barbara E Ainsworth. Associations of food consumption, serum vitamins and metabolic syndrome risk with physical activity level in middle-aged adults: the national health and nutrition examination survey (nhanes) 2005–2006. *Public health nutrition*, 19(9):1674–1683, 2016.

Anne HY Chu, Sheryl HX Ng, David Koh, and Falk Müller-Riemenschneider. Reliability and validity of the self-and interviewer-administered versions of the global physical activity questionnaire (gpaq). *PloS one*, 10(9):e0136944, 2015.

Dylan P Cliff, Kylie D Hesketh, Stewart A Vella, Trina Hinkley, Margarita D Tsiros, Nicola D Ridgers, Alison Carver, Jenny Veitch, A-M Parrish, Louise L Hardy, et al. Objectively measured sedentary behaviour and health and development in children and adolescents: systematic review and meta-analysis. *Obesity Reviews*, 17(4):330–344, 2016.

Ciprian M Crainiceanu and David Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185, 2004.

Dorothea Dumuid, Tyman E Stanford, Josep-Antoni Martin-Fernández, Željko Pedišić, Carol A Maher, Lucy K Lewis, Karel Hron, Peter T Katzmarzyk, Jean-Philippe Chaput, Mikael Fogelholm, et al. Compositional data analysis for physical activity, sedentary time and sleep research. *Statistical methods in medical research*, 27(12):3726–3738, 2018.

Dorothea Dumuid, Željko Pedišić, Tyman Everleigh Stanford, Josep-Antoni Martín-Fernández, Karel Hron, Carol A Maher, Lucy K Lewis, and Timothy Olds. The compositional isotemporal substitution model: a method for estimating changes in a health outcome for reallocation of time between sleep, physical activity and sedentary behaviour. *Statistical methods in medical research*, 28(3):846–857, 2019.

Juan José Egozcue and Vera Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37:795–828, 2005.

Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35 (3):279–300, 2003.

Juan José Egozcue, V Pawlowsky-Glahn, G Mateu-Figueras, and C Barceló-Vidal. Compositional data analysis: theory and applications. *Calculus of simplex-valued functions*, pages 158–175, 2011.

Pietro Ferrari, Christine Friedenreich, and Charles E Matthews. The role of measurement error in estimating levels of physical activity. *American journal of epidemiology*, 166(7):832–840, 2007.

William L Haskell, I-Min Lee, Russell R Pate, Kenneth E Powell, Steven N Blair, Barry A Franklin, Caroline A Macera, Gregory W Heath, Paul D Thompson, and Adrian Bauman. Physical activity and public health: updated recommendation for adults from the american college of sports medicine and the american heart association. *Circulation*, 116(9):1081, 2007.

Genevieve N Healy, David W Dunstan, Jo Salmon, Ester Cerin, Jonathan E Shaw, Paul Z Zimmet, and Neville Owen. Objectively measured light-intensity physical activity is independently associated with 2-h plasma glucose. *Diabetes care*, 30(6):1384–1389, 2007.

Genevieve N Healy, Charles E Matthews, David W Dunstan, Elisabeth AH Winkler, and Neville Owen. Sedentary time and cardio-metabolic biomarkers in us adults: Nhanes 2003–06. *European heart journal*, 32(5):590–597, 2011.

Max Hirshkowitz, Kaitlyn Whiton, Steven M Albert, Cathy Alessi, Oliviero Bruni, Lydia DonCarlos, Nancy Hazen, John Herman, Eliot S Katz, Leila Kheirandish-Gozal, et al. National sleep foundation's sleep time duration recommendations: methodology and results summary. *Sleep health*, 1(1):40–43, 2015.

James Honaker, Jonathan N Katz, and Gary King. A fast, easy, and efficient estimator for multiparty electoral data. *Political Analysis*, 10(1):84–100, 2002.

Karel Hron, Peter Filzmoser, and Katherine Thompson. Linear regression with compositional explanatory variables. *Journal of applied statistics*, 39(5):1115–1128, 2012.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Ian Janssen, Anna E Clarke, Valerie Carson, Jean-Philippe Chaput, Lora M Giangregorio, Michelle E Kho, Veronica J Poitras, Robert Ross, Travis J Saunders, Amanda Ross-White, et al. A systematic review of compositional data analysis studies examining associations between sleep, sedentary behaviour, and physical activity with health outcomes in adults. *Applied physiology, nutrition, and metabolism*, 45(10):S248–S257, 2020.

Dinesh John and Patty Freedson. Actigraph and actical physical activity monitors: a peek under the hood. *Medicine and science in sports and exercise*, 44(1 Suppl 1):S86, 2012.

Marta Karas, Jiawei Bai, Marcin Straczkiewicz, Jaroslaw Harezlak, Nancy W Glynn, Tamara Harris, Vadim Zipunnikov, Ciprian Crainiceanu, and Jacek K Urbanek. Accelerometry data in health

research: Challenges and opportunities: Review and examples. *Statistics in biosciences*, 11: 210–237, 2019.

Lisa Kobos, Christina R Ferreira, Tiago JP Sobreira, Bartek Rajwa, and Jonathan Shannahan. A novel experimental workflow to determine the impact of storage parameters on the mass spectrometric profiling and assessment of representative phosphatidylethanolamine lipids in mouse tissues. *Analytical and bioanalytical chemistry*, 413:1837–1849, 2021.

Catherine M Kotz and James A Levine. Role of nonexercise activity thermogenesis (neat) in obesity. *Minnesota medicine*, 88(9):54–57, 2005.

Andrea Z LaCroix, Eileen Rillamas-Sun, David Buchner, Kelly R Evenson, Chongzhi Di, I-Min Lee, Steve Marshall, Michael J LaMonte, Julie Hunt, Lesley Fels Tinker, et al. The objective physical activity and cardiovascular disease health in older women (opach) study. *BMC public health*, 17:1–12, 2017.

David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: a valid alternative to correlation for relative data. *PLoS computational biology*, 11(3):e1004075, 2015.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

Carol Maher, Tim Olds, Emily Mire, and Peter T Katzmarzyk. Reconsidering the sedentary behaviour paradigm. *PloS one*, 9(1):e86403, 2014.

Barceló-Vidal Martín-Fernández. Pawlowsky-glahn (2003) martín-fernández ja, barceló-vidal c, pawlowsky-glahn v. dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003.

Rania A Mekary, Walter C Willett, Frank B Hu, and Eric L Ding. Isotemporal substitution paradigm for physical activity epidemiology and weight change. *American journal of epidemiology*, 170(4):519–527, 2009.

Rania A Mekary, Michel Lucas, An Pan, Olivia I Okereke, Walter C Willett, Frank B Hu, and Eric L Ding. Isotemporal substitution analysis for physical activity, television watching, and risk of depression. *American journal of epidemiology*, 178(3):474–483, 2013.

Jean Carlos Montero-Serrano, Javier Palarea-Albaladejo, Josep A Martín-Fernández, Manuel Martínez-Santana, and José Vicente Gutiérrez-Martín. Sedimentary chemofacies characterization by means of multivariate analysis. *Sedimentary Geology*, 228(3-4):218–228, 2010.

National Center for Health Statistics. NHANES 2005-2006 data. https://wwwn.cdc.gov/nchs/nhanes/search/DataPage.aspx?Component=Questionnaire&CycleBeginYear=2005, 2005.

Heather K Neilson, Paula J Robson, Christine M Friedenreich, and Ilona Csizmadi. Estimating activity energy expenditure: how valid are physical activity questionnaires? *The American journal of clinical nutrition*, 87(2):279–291, 2008.

NHANES. National health and nutrition examination survey. https://www.cdc.gov/nchs/nhanes/index.htm.

Australian Government Department of Health. Australia's physical activity and sedentary behaviour guidelines. *Dep Heal Website*, 2014.

Chief Medical Officer. Start active, stay active: a report on physical activity from the four home countries' chief medical officers. 2013.

Neville Owen, Adrian Bauman, and Wendy Brown. Too much sitting: a novel and important predictor of chronic disease risk? *British journal of sports medicine*, 43(2):81–83, 2009.

Javier Palarea-Albaladejo and Josep-Antoni Martín-Fernández. A modified em alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34(8):902–917, 2008.

Javier Palarea-Albaladejo, Josep A Martín-Fernández, and Juan Gómez-García. A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39:625–645, 2007.

Karl Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, 60(359-367):489–498, 1897.

Željko Pedišić. Measurement issues and poor adjustments for physical activity and sleep undermine sedentary behaviour research—the focus should shift to the balance between sleep, sedentary behaviour, standing and activity. *Kinesiology*, 46(1.):135–146, 2014.

Željko Pedišić and Adrian Bauman. Accelerometer-based measures in physical activity surveillance: current practices and issues. *British journal of sports medicine*, 49(4):219–223, 2015.

Michele ER Pierotti, Josep A Martín-Fernández, and Ole Seehausen. Mapping individual variation in male mating preference space: multiple choice in a color polymorphic cichlid fish. *Evolution*, 63(9):2372–2388, 2009.

Veronica Joan Poitras, Casey Ellen Gray, Michael M Borghese, Valerie Carson, Jean-Philippe Chaput, Ian Janssen, Peter T Katzmarzyk, Russell R Pate, Sarah Connor Gorber, Michelle E Kho, et al. Systematic review of the relationships between objectively measured physical activity and health indicators in school-aged children and youth. *Applied physiology, nutrition, and metabolism*, 41(6):S197–S239, 2016.

William S Rayens and Cidambi Srinivasan. Box–cox transformations in the analysis of compositional data. *Journal of Chemometrics*, 5(3):227–239, 1991a.

William S Rayens and Cidambi Srinivasan. Estimation in compositional data analysis. *Journal of chemometrics*, 5(4):361–374, 1991b.

Miriam Reiner, Christina Niermann, Darko Jekauc, and Alexander Woll. Long-term health benefits of physical activity–a systematic review of longitudinal studies. *BMC public health*, 13(1):1–9, 2013.

Roger Ros-Freixedes and Joan Estany. On the compositional analysis of fatty acids in pork. *Journal of agricultural, biological, and environmental statistics*, 19:136–155, 2014.

Mary E Rosenberger, Janet E Fulton, Matthew P Buman, Richard P Troiano, Michael A Grandner, David M Buchner, and William L Haskell. The 24-hour activity cycle: a new paradigm for physical activity. *Medicine and science in sports and exercise*, 51(3):454, 2019.

JL Scealy and AH2815780 Welsh. Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):351–375, 2011.

JL Scealy and Alan H Welsh. Fitting kent models to compositional data with small concentration. *Statistics and Computing*, 24:165–179, 2014.

JL Scealy, Patrice de Caritat, Eric C Grunsky, Michail T Tsagris, and AH Welsh. Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association*, 110(509):136–148, 2015.

Jennifer A Schrack, Vadim Zipunnikov, Jeff Goldsmith, Jiawei Bai, Eleanor M Simonsick, Ciprian Crainiceanu, and Luigi Ferrucci. Assessing the "physical cliff": detailed quantification of age-related differences in daily patterns of physical activity. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 69(8):973–979, 2014.

Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

Stephanie Skender, Jennifer Ose, Jenny Chang-Claude, Michael Paskow, Boris Brühmann, Erin M Siegel, Karen Steindorf, and Cornelia M Ulrich. Accelerometry and physical activity questionnaires-a systematic review. *BMC public health*, 16(1):1–10, 2016.

Connie Stewart and Christopher Field. Managing the essential zeros in quantitative fatty acid signature analysis. *Journal of Agricultural, Biological & Environmental Statistics (JABES)*, 16 (1), 2011.

Robert A Stine. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.

Mark S Tremblay, Allana G LeBlanc, Michelle E Kho, Travis J Saunders, Richard Larouche, Rachel C Colley, Gary Goldfield, and Sarah Connor Gorber. Systematic review of sedentary behaviour and health indicators in school-aged children and youth. *International journal of behavioral nutrition and physical activity*, 8(1):1–22, 2011.

Mark S Tremblay, Valerie Carson, Jean-Philippe Chaput, Sarah Connor Gorber, Thy Dinh, Mary Duggan, Guy Faulkner, Casey E Gray, Reut Gruber, Katherine Janson, et al. Canadian 24-hour movement guidelines for children and youth: an integration of physical activity, sedentary behaviour, and sleep. *Applied physiology, nutrition, and metabolism*, 41(6):S311–S327, 2016.

Michail Tsagris. Regression analysis with compositional data containing zero values. *arXiv preprint arXiv:1508.01913*, 2015.

Michail Tsagris and Connie Stewart. A folded model for compositional data analysis. *Australian & New Zealand Journal of Statistics*, 62(2):249–277, 2020.

Michail Tsagris, Simon Preston, and Andrew TA Wood. Improved classification for compositional data using the α-transformation. *Journal of classification*, 33:243–261, 2016.

Michail T Tsagris, Simon Preston, and Andrew TA Wood. A data-based power transformation for compositional data. *arXiv preprint arXiv:1106.1451*, 2011.

Matthew CB Tsilimigras and Anthony A Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology*, 26(5):330–335, 2016.

Jacek K Urbanek, Adam P Spira, Junrui Di, Andrew Leroux, Ciprian Crainiceanu, and Vadim Zipunnikov. Epidemiology of objectively measured bedtime and chronotype in us adolescents and adults: Nhanes 2003–2006. *Chronobiology international*, 35(3):416–434, 2018.

Jeff K Vallance, Elisabeth AH Winkler, Paul A Gardiner, Genevieve N Healy, Brigid M Lynch, and Neville Owen. Associations of objectively-assessed physical activity and sedentary time with depression: Nhanes (2005–2006). *Preventive medicine*, 53(4-5):284–288, 2011.

Vincent T Van Hees, Lukas Gorzelniak, Emmanuel Carlos Dean León, Martin Eder, Marcelo Pias, Salman Taherian, Ulf Ekelund, Frida Renström, Paul W Franks, Alexander Horsch, et al. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PloS one*, 8(4):e61691, 2013.

Vijay R Varma, Debangan Dey, Andrew Leroux, Junrui Di, Jacek Urbanek, Luo Xiao, and Vadim Zipunnikov. Re-evaluating the effect of age on physical activity over the lifespan. *Preventive medicine*, 101:102–108, 2017.

Huiwen Wang, Qiang Liu, Henry MK Mok, Linghui Fu, and Wai Man Tse. A hyperspherical transformation forecasting model for compositional data. *European journal of operational research*, 179(2):459–468, 2007.

Richard A Washburn, Alan M Jette, and Carol A Janney. Using age-neutral physical activity questionnaires in research with the elderly. *Journal of Aging and Health*, 2(3):341–356, 1990.

Duan Weipeng, Garry Kuan, Lou Hu, and Yee Cheng Kueh. Development and prospect of isotemporal substitution model in physical activity research: A narrative review. In *Advancing Sports and Exercise via Innovation: Proceedings of the 9th Asian South Pacific Association of Sport Psychology International Congress (ASPASP) 2022, Kuching, Malaysia*, pages 335–353. Springer, 2023.

Klaas R Westerterp. Physical activity assessment with accelerometers. *International Journal of Obesity*, 23(3):S45–S49, 1999.

Emma G Wilmot, Charlotte L Edwardson, Felix A Achana, Melanie J Davies, Trish Gorely, Laura J Gray, Kamlesh Khunti, Thomas Yates, and Stuart JH Biddle. Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. *Diabetologia*, 55(11):2895–2905, 2012.

t World Health Organization et al. *Global recommendations on physical activity for health*. World Health Organization, 2010.

Bo Xi, Dan He, Min Zhang, Jian Xue, and Donghao Zhou. Short sleep duration predicts risk of metabolic syndrome: a systematic review and meta-analysis. *Sleep medicine reviews*, 18(4): 293–297, 2014.

Luo Xiao, Lei Huang, Jennifer A Schrack, Luigi Ferrucci, Vadim Zipunnikov, and Ciprian M Crainiceanu. Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics*, 16(2):352–367, 2015.

Che-Chang Yang and Yeh-Liang Hsu. A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors*, 10(8):7772–7788, 2010.

Waleed A Yousef. Estimating the standard error of cross-validation-based estimators of classifier performance. *Pattern Recognition Letters*, 146:115–125, 2021.

G Zadora and T Neocleous. Likelihood ratio model for classification of forensic evidence. *Analytica Chimica Acta*, 642(1-2):266–278, 2009.